Introduction to NLG

Ehud Reiter

Who is Ehud

- Professor of Computing Science at University of Aberdeen
- Formerly Chief Scientist of Arria NLG (spinout I co-founded)
- Working on NLG and Evaluation since 1990s

Goal

- Help you do a better job evaluating language generation systems
  - Focus on basics, what goes wrong
  - Focus on experimental design more than technology
- Help you identify untrustworthy evaluations
- Give you some general background on NLG

Ehud’s Book

- Much of the material I am presenting comes from a new book on NLG which I have written
  - Should be published in October or November
- Best source for going deeper into material

E Reiter (2024). *Natural Language Generation*. Springer

Natural Language Generation

- AI systems that generate texts in English and other human lang
- I will focus on systems that generate output from inputs
  - Sports data -> sports story
  - Driving data -> safety advice
  - Doctor-patient consultation -> summary for patient record
- *Evaluation*
  - How can we tell if generated texts are high-quality and appropriate?
  - Focus is on helping people do good evaluations, not on evaluation research.

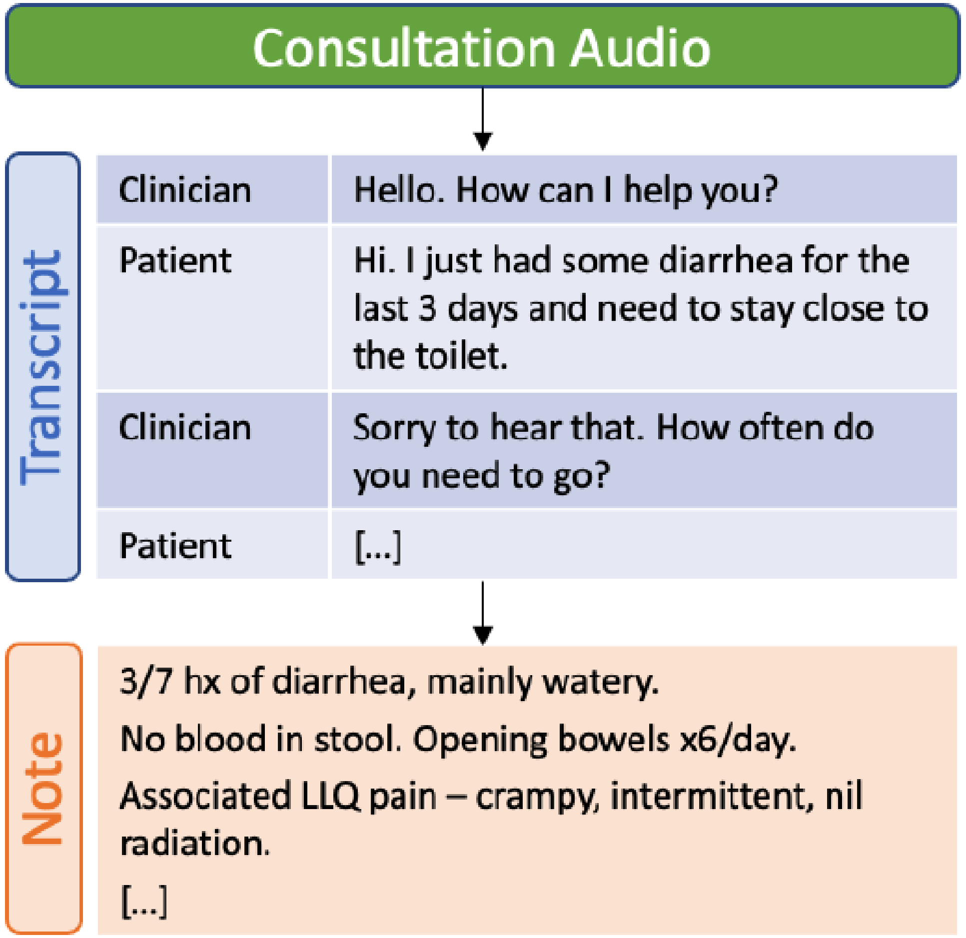
What is NLG?

- *Example NLG systems*
- Technology: rules, ML, LLM
- Applications
- History

Example: Summarise Consultations

- Input: doctor-patient consultation
- Output: summary of consultation for patient record
- Many other NLG reporting applications

T Knoll et al (2022). User-Driven Research of Medical Note Generation Software. *Proc of NAACL*



Example: Sportswriter

- Input is statistics about a basketball game
- Output is sports text for fans or bettors
- Many other media apps

C Thomson, et al (2023). Evaluating factual accuracy in complex data-to-text. *Computer Speech and Language*

TEAM	W	L	H1-PTS	H2-PTS	PTS	FG%
Grizzlies	5	0	46	56	102	.486
Suns	3	2	52	39	91	.559

Player	TEAM	PTS	REB	AST	BLK	STL
Marc Gasol	Grizzlies	18	5	6	0	4
Isaiah Thomas	Suns	15	1	2	0	1

The Memphis Grizzlies (5-2) defeated the Phoenix Suns (3-2) Monday 102-91 at the Talking Stick Resort Arena in Phoenix. The Grizzlies had a strong first half where they out-scored the Suns 59-42. Marc Gasol scored 18 points, leading the Grizzlies. Isaiah Thomas added 15 points, he is averaging 19 points on the season so far.

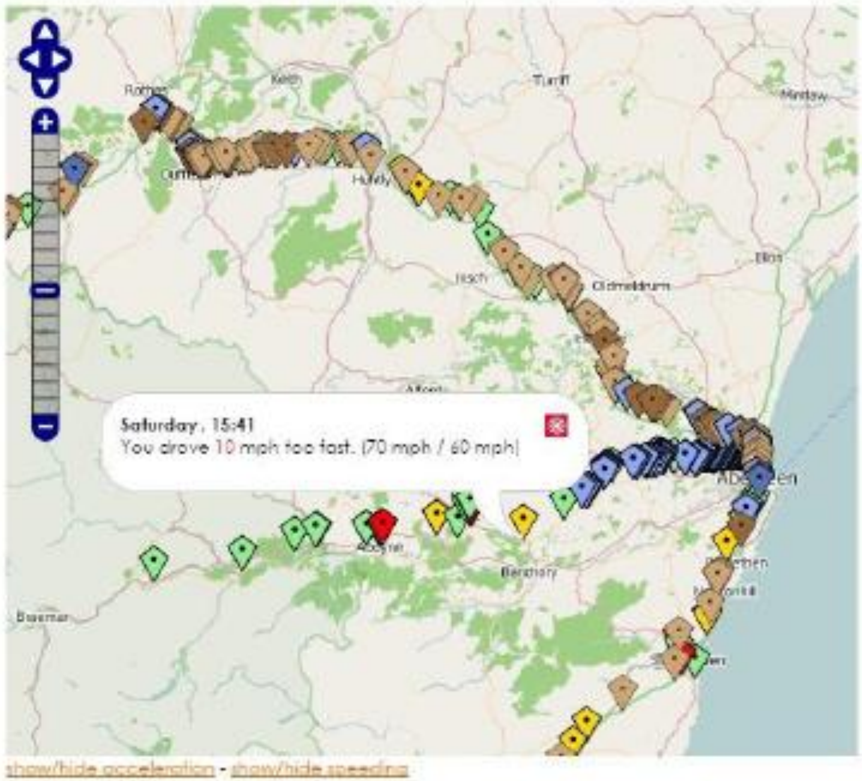
Example: Driving Report

- Input is GPS data which records cars position, speed, etc over a week
- Output is feedback report on unsafe driving
- I work on many apps to summarise complex personal data, esp in health

D Braun, et al (2018). SaferDrive: An NLG-based behaviour change support systems for drivers. *Natural Language Engineering*

Driving Report 31 August - 6 September

You drove **170 kilometres in five hours and 40 minutes** during the last week. You managed to reduce the number of acceleration incidents per kilometre by **nearly 10%**, well done!



You didn't do any serious speeding, well done! However, you **speeded on** 31 occasions, 15 times **on weekdays on King Street**.

You **accelerated or braked harshly** 103 times, mostly on weekdays on A91 and on weekdays on roads with 30 mph speed limit in Aberdeen.

Many more applications!

- In my career, I have also worked on NLG for
  - Summarising medical records for doctors, nurses, patients
  - Explaining educational assessment results to students
  - Helping maintenance engineers for offshore oil rigs find problems
  - Generating stop-smoking advice
  - Supporting non-verbal children in communicating
- Loads of other use cases!

Technology

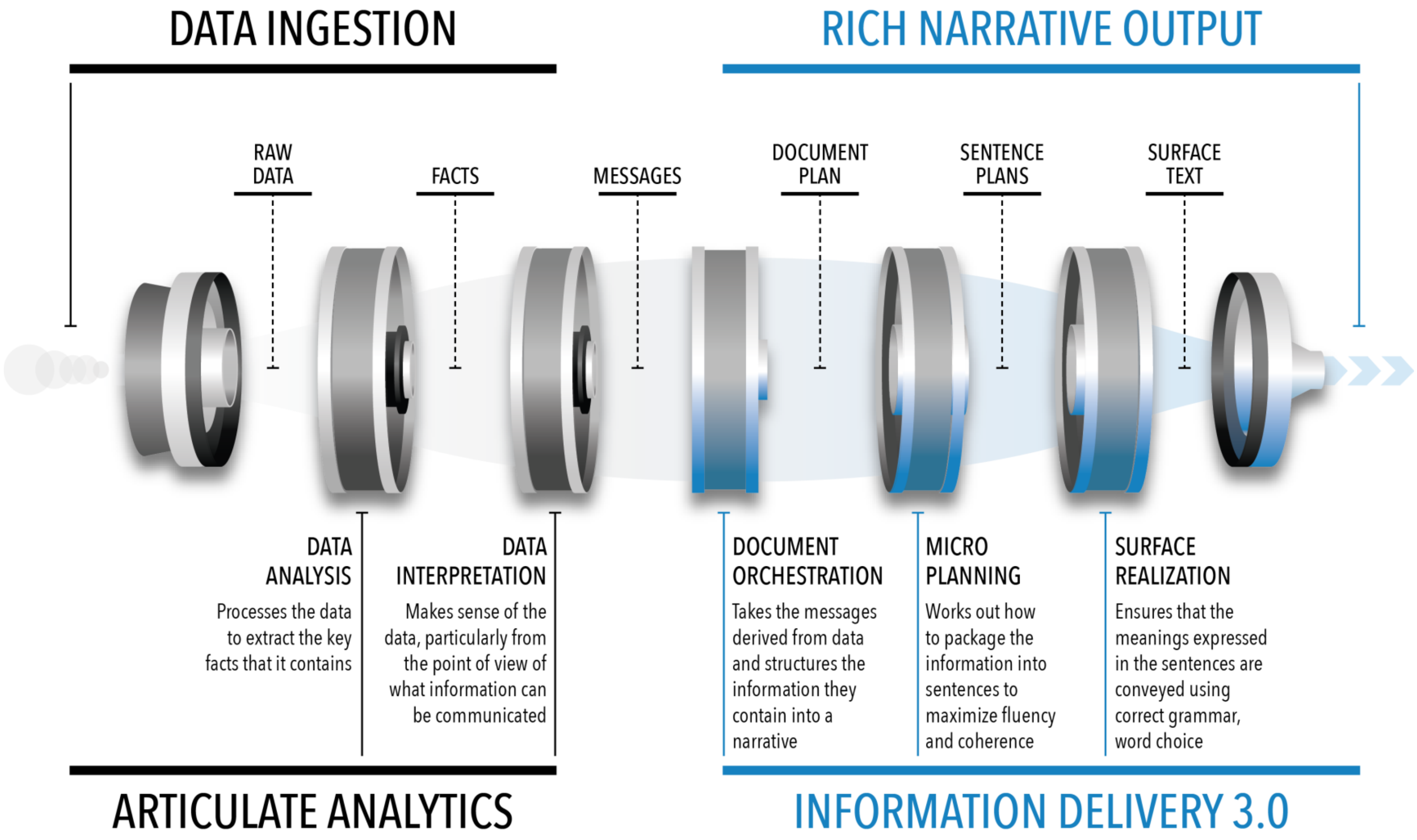
- *Rules*: Write explicit rules and algorithms to generate texts
- *Fine-tune (or train) models*: Fine-tune a language model to generate texts in a domain
- *Prompt model*: Prompt a language model to generate a text

Rule-based NLG

- Write rules and algorithms to generate texts
- A lot of work, but gives full control over system
  - Bugs are relatively easy to fix
- Still popular in many mission/safety critical applications
- Understanding rule-based perspective can help when using other technologies

- Often done using a generation *pipeline*

Pipeline (Arria)



Pipeline stages

- *Input Data*
- *Signal analysis*: Find potentially important patterns in data
- *Data interpretation*: Create useful insights about the data
- *Document planning*: Choose content (insights mentioned) and structure (sentences and paragraphs) of text
- *Microplanning*: Decide how to linguistically express insights
- *Surface realisation*: Generate grammatically correct texts

E Reiter (2007). An Architecture for Data-to-Text Systems. *Proc of ENLG-2007*

Example: driving feedback

Input: Position and speed from GPS; street name and speed limit from gazeteer lookup on position

time	*speed*	*street*
9:00:00	30	King Street
9:00:01	32	King Street
9:00:02	35	King Street
9:00:03	30	King Street
9:00:04	33	King Street
9:00:05	32	King Street
9:00:06	27	King Street
9:00:07	25	King Street
9:00:08	30	King Street
9:00:09	33	King Street
9:00:10	33	King Street
9:00:11	30	King Street
9:00:12	25	King Street
9:00:13	20	King Street

Speed limit on King Street is 30mph

Signal analysis

Signal analysis: find potentially important patterns

- Mostly use standard pattern detection techniques
- Noise detection/suppression often important

Time periods where speeding occurred:

Speeding(9:00:01-9:00:05, King Street, maxSpeed=35)
Speeding(9:00:09-9:00:10, King Street, maxSpeed=33)

Data Interpretation

Data interpretation: Create useful insights about the data

- Domain-specific analytics
- Rules from domain experts

Clusters of related speeding segments

SpeedingCluster(9:00:01-9:00:10, King Street, maxSpeed=35, incidents=2)

Document Planning

Document Planning: Choose content (insights mentioned) and structure (sentences and paragraphs) of text

Document plan:

Sentence: SpeedingCluster(9:00:01-9:00:10, King Street, maxSpeed=35, incidents=2)

Microplanning

Microplanning: Decide how to linguistically express insights

Choose words and syntactic structures

Insight: SpeedingCluster(9:00:01-9:00:10, King Street, maxSpeed=35, incidents=2)

Sentence: Subject=you, verb=speed (past tense), modifier=twice, location=on King Street

Surface Realisation

Surface realisation: Generate grammatically correct texts

Final sentence

You sped twice on King Street.

Rule-based NLG pipeline

- Some open-source tools and libraries to support this
  - Simplenlg
- Also proprietary commercial tools
- Many research papers on individual stages
  - My new book is good source

Discussion

- Companies in professional sectors sometimes tell me “we tried LLMs, but decided rule-based was better for us”
- Why would they say this?

Neural models

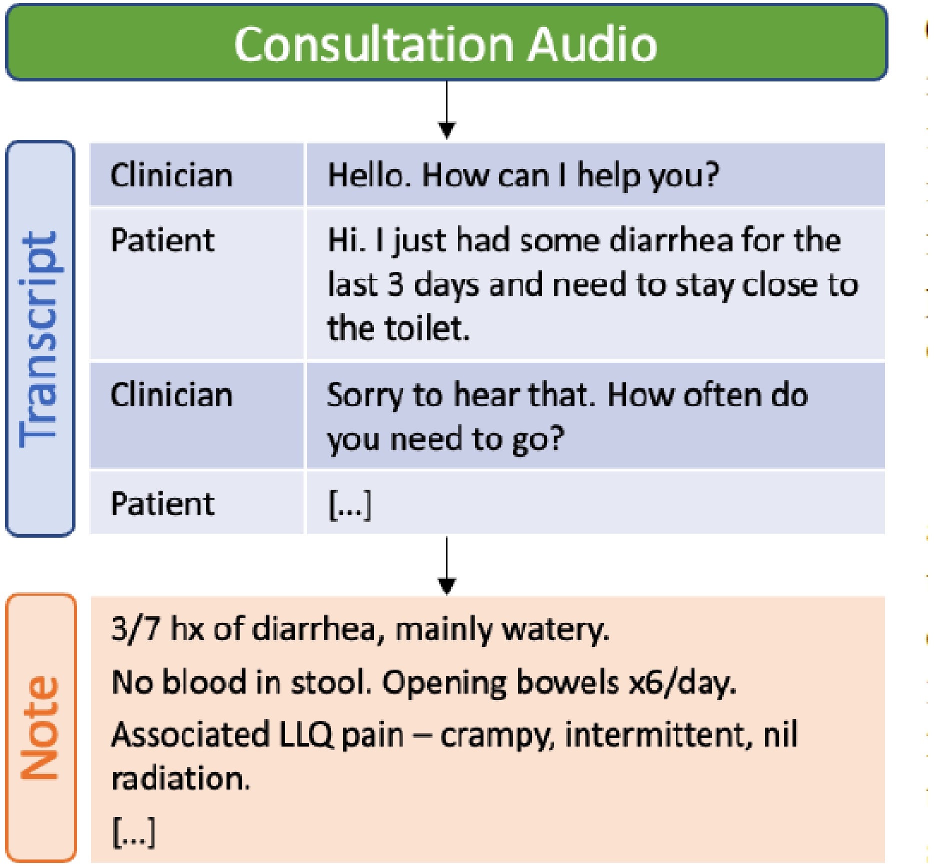
- Build model from data (eg, LSTM) – not used, will not discuss
- Fine-tune language model (eg, BART)
- Prompt LLM (eg, GPT4)

Fine-tune language model

- Start with general-purpose large language model
  - Can use GPT/Llama/etc, but people also still use older models such as BERT and BART
- Fine-tune it on domain corpus

Example: Summarising Consultations

NoteGenerator system built by fine-tuning BART model on 10000 real-world consultations and associated human-writren summaries



Data issues

- Wont discuss technology here
  - Changing very fast
  - Most people know something about it
- Look at data issues
  - Important, change less over time
  - Good data more important than using latest model

Training data

- Need significant amounts of training data to build high-quality NLG system
  - Typically 1000s of input-output pairs
- Must be of high-quality
  - Be wary of Kaggle data sets!
  - Companies spend significant sums on data quality
- Must be representative and have good coverage
  - Don’t use dialogues from TV shows to train dialogue systems for real people
  - Don’t assume good performance on data from one hospital means good performance on all hospitals

Data issues

- Domain shift – the world changes, so models trained on 5-year old data may not work well.
  - Medicine: new interventions, regulations, equipment
  - Law: laws change!
- Edge cases – Even 10000 input-output pars will not cover unusual edge or boundary cases
  - Safety issues

Workflow

- Sometimes models used without human supervision
- Other cases have “human-in-loop” where a person checks and edits the models output.
- NoteGenerator – doctors checked and edited summaries
  - In complex or unusual cases, often threw out summaries and wrore from scratch
- Models can also be used to check human-written output, and to generate material which humans can copy-and-paste
- Need good UI design and workflow integration!

Prompted language model

- Simply ask GPT/Llama/etc to do something via a prompt
  - Input data included in prompt

Generate a short weather forecast from the below data.
Weather data:
[{"time": 0, "wind": 10, "precipitation": 0, "temperature": 6},
{"time": 3, "wind": 11, "precipitation": 0, "temperature": 6},
…]

Output of prompted model:
Here’s a short weather forecast based on the provided data:
Expect a day with changing wind speeds and temperatures. The wind will range from 9 km/h to 12 km/h throughout the day, occasionally gusting higher. The temperature will start at 6°C and gradually increase to 15°C in the afternoon before dropping to 10°C in the evening. There is no precipitation expected, so it will be a dry day. In summary, anticipate a mix of wind speeds, with temperatures ranging from 6°C to 15°C, and no rainfall.

Prompted LLMs

- No training data needed!
  - A few examples can be provided in prompt, this is optional
  - Huge benefit
- Texts benefit from domain knowledge inside LLM
  - Background, causality, etc
- Naturally support dialogue, so users can ask followup questions
- Output texts almost always read well
- Inherent support for lots of languages

Data issues

- Prompted LLMs usually trained on Internet data
  - Only thing large enough
- Dangers
  - Low-quality or inappropriate content: Internet full of this
  - Malicious or spam content: will hackers try to get this into LLMs
  - Out-of-date content: leads to incorrect texts
  - Biased or racist content: lots of this around
- Lack of control
  - Hard to tweak systems or fix bugs
  - Hard to test

Hallucinations

- LLMs can produce incorrect and even dangerous texts
- GPT4 advice on support for cancer patients in Aberdeen

> CLAN Cancer Support: They offer a range of services including counselling, relaxation therapies, and support groups.
>
> - Address: CLAN House, 120 Westburn Road, Aberdeen, AB25 2QA.
> - Website: [CLAN Cancer Support](http://www.clanhouse.org)

- CLAN is great, but URL is for a spam site

M Sun et al (2024). Effectiveness of ChatGPT in explaining complex medical reports to patients. Arxiv

Dealing with hallucinations

- Choose use case where doesn’t matter much
  - Sportswriting (human texts often wrong)
- Human-in-loop workflow
- Separate checking module
- Better technology
  - Better tech has reduced problem, but by no means eliminated it
  - Inherent to LLMs?

Discussion

- Which approach would you use for
  - AI copywriter (marketing material)
  - AI sportswriter
  - AI tool gives financial advice to investors (conforming to corporate brand)
  - AI app tells people when they need to contact their doctor

What is NLG?

- Example NLG systems
- Technology: rules, ML, LLM
- *Applications*
- History

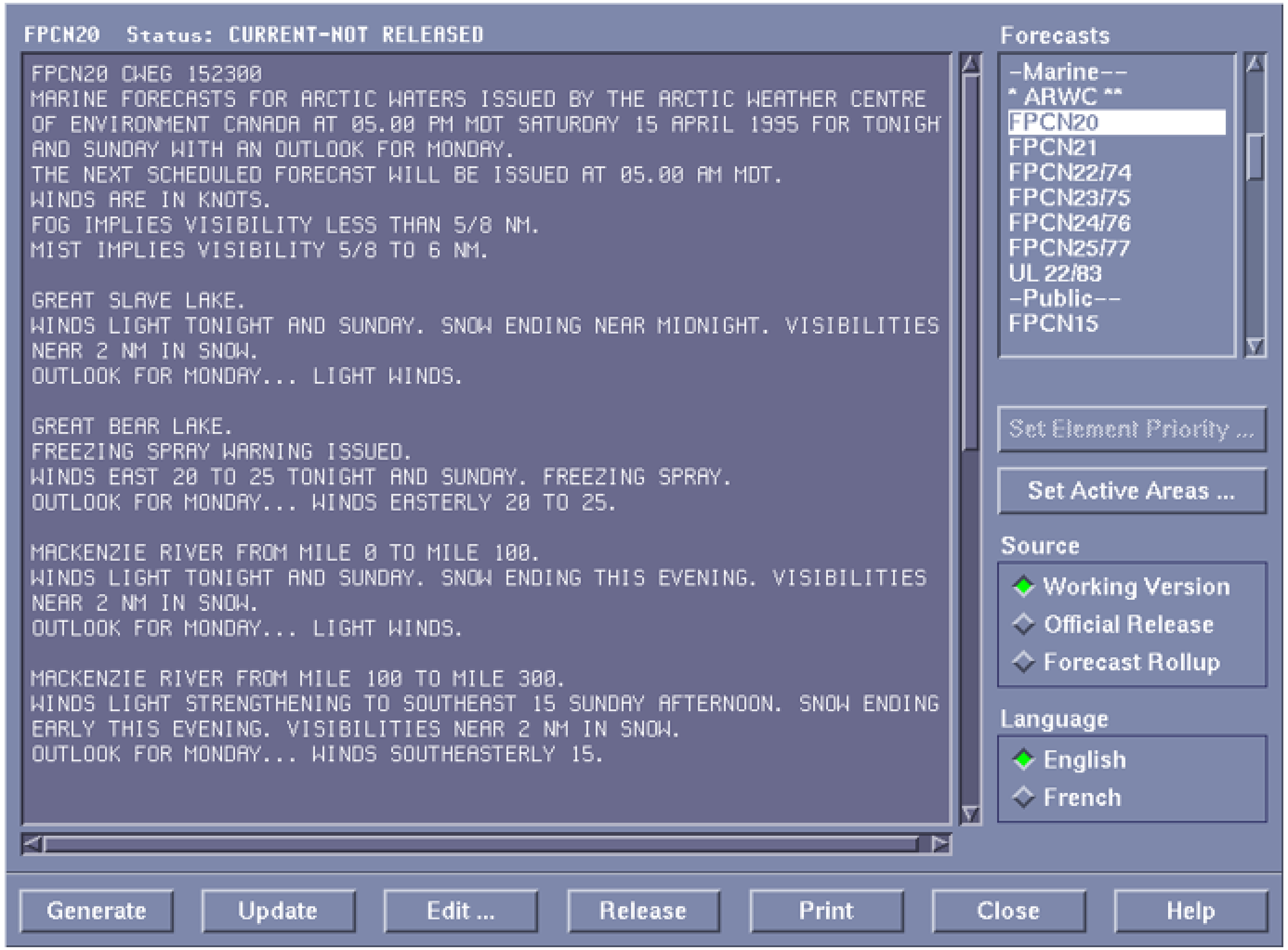
Long-standing applications of NLG

- Developers are experimenting with NLG in lots of different areas
- Focus here on area where NLG has been used for many years

Weather forecasts

- First real-world use for NLG
- First production went live in 1992

E Goldberg et al (1994). Using natural-language processing to produce weather forecasts. *IEEE Expert*



Automatic Journalism

- NLG used to generate draft news stories for past 10 years
  - Esp local news
- Checked by journalist

https://ehudreiter.com/2019/12/23/election-results-lessons-from-a-real-world-nlg-system/

Florence Eshalomi has been elected MP for Vauxhall, meaning that the Labour Party holds the seat with a decreased majority.

The new MP beat Liberal Democrat Sarah Lewis by 19,612 votes. This was fewer than Kate Hoey's 20,250-vote majority in the 2017 general election.

Sarah Bool of the Conservative Party came third and the Green Party's Jacqueline Bond came fourth.

Voter turnout was down by 3.5 percentage points since the last general election.

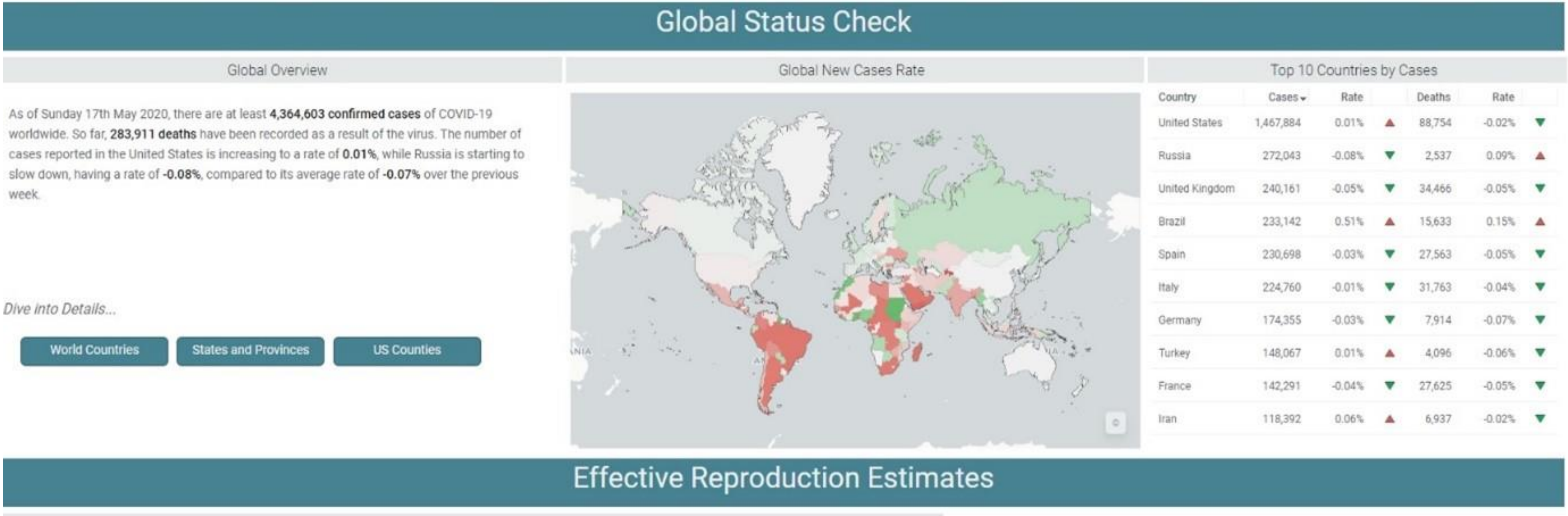
More than 56,000 people, 63.5\% of those eligible to vote, went to polling stations across the area on Thursday, in the first December general election since 1923.

etc

Business Intelligence

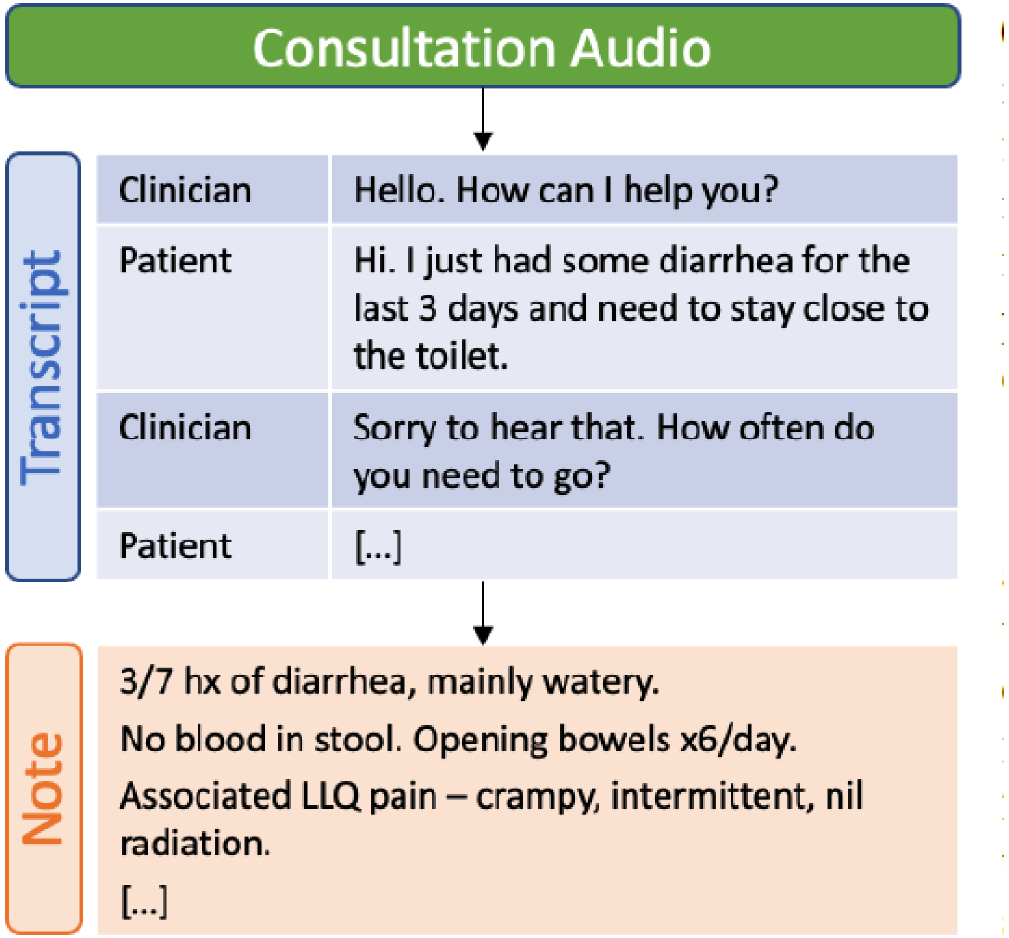
NLG widely used to generate text to support business intelligence graphics

https://ehudreiter.com/2020/05/21/adding-narrative-to-a-covid-dashboard/



Summarisation

Summarise documents or conversations



Medical

- Reporting
- Patient information
- Behaviour change
- Decision support
- Etc

Smoking Information for Heather Stewart

You have good reasons to stop...

People stop smoking when they really want to stop. It is encouraging that you have many good reasons for stopping. The scales show the good and bad things about smoking for you. They are tipped in your favour.

THINGS YOU LIKE
it's relaxing
it stops stress
you enjoy it
it relieves boredom
it stops weight gain
it stops you craving

THINGS YOU DISLIKE
it makes you less fit
it's a bad example for kids
you're addicted
it's unpleasant for others
other people disapprove
it's a smelly habit
it's bad for you
it's expensive
it's bad for others' health

You could do it...

Most people who really want to stop eventually succeed. In fact, 10 million people in Britain have stopped smoking - and stayed stopped - in the last 15 years. Many of them found it much easier than they expected.

Although you don't feel confident that you would be able to stop if you were to try, you have several things in your favour.

- You have stopped before for more than a month.
- You have good reasons for stopping smoking.
- You expect support from your family, your friends, and your workmates.

We know that all of these make it more likely that you will be able to stop. Most people who stop smoking for good have more than one attempt.

Overcoming your barriers to stopping...

You said in your questionnaire that you might find it difficult to stop because smoking helps you cope with *stress*. Many people think that cigarettes help them cope with stress. However, taking a cigarette only makes you feel better for a short while. Most ex-smokers feel calmer and more in control than they did when they were smoking. There are some ideas about coping with stress on the back page of this leaflet.

You also said that you might find it difficult to stop because you would *put on weight*. A few people do put on some weight. If you did stop smoking, your appetite would improve and you would taste your food much better. Because of this it would be wise to plan in advance so that you're not reaching for the biscuit tin all the time. Remember that putting on weight is an overeating problem, not a no-smoking one. You can tackle it later with diet and exercise.

And finally...

We hope this letter will help you feel more confident about giving up cigarettes. If you have a go, you have a real chance of succeeding.

With best wishes,

The Health Centre.

What is NLG?

- Example NLG systems
- Technology: rules, ML, LLM
- Applications
- *History*

Up to 2010

- 1600s: speculation about mechanical generation of text
- 1960s: computer generation of texts within machine translation
- 1970s: first non-MT NLG research
- 1980s: specialist NLG conferences and interest groups
- 1990s: first production NLG applications
- 2000s: evaluation taken much more seriously

Recent history

2010-2020

- Rapidly rising number of commercial NLG companies (inc Arria)
- Neural generation techniques
- Shared tasks

2020-now

- Too recent to describe as “history”