Als Machines Liegen / When Machines Lie

Kees van Deemter

c.j.vandeemter@uu.nl

Utrecht, 21 January 2020

A slightly enhanced version of the slides used at my inaugural lecture of 12 Sept. 2019
Plan of this lecture

1. What we will mean by “lying” today
2. Lying in politics and in literature
3. Natural Language Generation (NLG)
4. Lessons from a famous NLG system:
   - “Lying” is hard to avoid
   - Effective communication requires “lying”
5. Benign decept (i.e., white lies)
6. Conclusion
A slight abuse of words

“Lying” = any deviation from the truth and the whole truth

– On purpose or by accident
– Incorrect or incomplete information
(Linguists: the Maxims of Grice!)

I’ll omit the scare quotes around “lying”
Background 1: Ian McEwan 2019

*Machines Like Me*

A novel about a human-like robot. Lying is depicted as hard to do.

“Who’s going to write the algorithm for the little white lie? (...) We do not yet know how to teach machines to lie”

(p.303)
Background 2

Fake News / Disinformation

Deliberate creation and sharing of false information
  – E.g., the Obama birther myth

Disinformation is not new

“The first casualty when war comes is truth” (Hiram W. Johnson 1918)

But social media & algorithms can boost dissemination

Facebook, Google, Weibo need to (be seen to) act:
  – E.g., Automatic fact checking
• Plenty of research on lying in many areas
  – Linguistics; Philosophy; Psychology; Logic; Game Theory; Law; ...

• Today’s talk offers an engineering perspective
  – Language technology/engineering
  – Specifically, Natural Language Generation (NLG)
This talk has 2 aims

• Introduce Natural Language Generation (NLG)
• Offer some insight into lying

From joint work with Ehud Reiter, Aberdeen
Something about Natural Language Generation (NLG)

Computer-generated texts, e.g.,
  – weather reports (Goldberg et al. 1994)
  – robot journalism (Theune et al. 2001, Latar 2014)
  – medical shift reports (Portet et al. 2009)

Systems that convert data into text
  – Often the text is only a summary
  – Have to select what’s important!
Natural Language Generation (NLG) systems

Growing interest from academia and industry

– Forbes Magazine (Jan. 2017): NLG one of “the 10 hottest AI technologies”

Imagine you were a sandal wearer ...
You (1970-present)
Recently, all around you
Example NLG system: Babytalk

EPSRC project Edinburgh/Aberdeen (e.g., Reiter)

**Sensors** produce numerical information, e.g.,
- Monitoring babies in intensive care
- Temperature, Heart Rate, ..., etc., 24/7
Baby Monitoring through sensors

- SpO2 (SO, HS)
- ECG (HR)
- Pericardial Temperature (TP)
- Arterial Line (Blood Pressure)
- Transcutaneous Probe (CO, OX)
- Core Temperature (TC)
Clinicians struggle to handle so much data

Babytalk generates text summaries automatically

– for clinicians: **BTnurse**
– for family: **BTclan**
Graphical rendering of 45 min interval
Neonatal ICU
Architecture of Babytalk

Four components arranged in a pipeline:

1. Signal Analysis
2. Data Interpretation
3. Content Determination
4. “Expressing it in words”
A report written by a nurse

At 1046 the baby is turned for re-intubation and re-intubation is complete by 1100 the baby being bagged with 60% oxygen between tubes. During the re-intubation there have been some significant bradycardias down to 60/min, but the sats have remained OK. The mean BP has varied between 23 and 56, but has now settled at 30. The central temperature has fallen to 36.1 °C and the peripheral temperature to 33.7 °C. The baby has needed up to 80% oxygen to keep the sats up.

Over the next 10 mins the HR decreases to 140 and the mean BP = 30-40. The sats fall with ETT suction so the FiO2 is increased to 80% but by 1112 the FiO2 is down to 49%.
By 11:00 the baby had been hand-bagged a number of times causing 2 successive bradycardias. She was successfully re-intubated after 2 attempts. The baby was sucked out twice. At 11:02 FIO2 was raised to 79%.
Limitation of this talk

We discuss **WHAT** an NLG system does

**Not** the techniques used (**HOW**)

**Not** how well these systems perform
1. Signal Analysis

A collection of mathematical tools

• Detect trends, patterns and events
  – Blood oxygen levels increasing
  – Upward spike in heart rate

• Infer missing values

• Change incorrect values
Example of incorrect values
Incorrect values adjusted
Where can signal analysis go wrong?

Inference of values is error prone

– 20 min → A spike in HR may be missed!
– Abnormality may be overlooked
  → Texts suggests all is well with the baby
  → Lying by omission
– Shows why we defined “lying” so liberally
  (the truth and the whole truth)
Linking patterns causally, e.g.,

“Heart Rate went up because baby was handled”

• Potentially a strong point of text!
  – Standard graphs do not show causal links
  – Human readers need a “story”

• But causal statement may be incorrect!
Simple example of a causation rule

*typical rule-based AI (Portet et al. 2009)*

CAUSES

( [HEART-RATE, [direction = increasing],
   EVENT, [type = HANDLING], 100 )

"If a handling event is followed by an increase in heart rate within 100 seconds, then the handling has caused the increase"

But this could easily be wrong!

Increased HR may have a deeper cause (e.g. infection)

This is not some fluke that is easy to fix

Causality remains ill-understood and difficult to prove
4. Expressing it in words
Some of the main parts

1. **Text Structuring:** How to organize the text? *Chronologically? Thematically?*

2. **Lexical Choice:** What words to use? *E.g.*, Use *medical jargon*? “**Bradycardia**” or “**Low heart rate**”? *(These expressions have roughly the same meaning.)*

3. **Referring Expressions Generation:** What reference strategy? *(e.g., van Deemter 2016, MIT Press)*
   *E.g.*, “**At 10:59 AM**”, or “**After handling**”? *(These expressions may refer to precisely the same time.)*
Referring Expressions Generation

Road gritters in trucks apply salt to slippery roads (Turner et al. 2009)

System says which parts of which roads will be slippery using speech (no graphics!)

Typically a long list (~ 1000 road stretches)

How to communicate this list to the drivers?
Road fragments that need salt
from weather predictions
Referring Expressions Generation

A literal description (enumeration):

- “Road $x_1$ and (...) and road $x_{1000}$”

  True but **verbose** $\rightarrow$ Not feasible

Rough approximations are computed, e.g.:

- “Areas above 200 metres”

  Statement is succinct but **not true**
  It has false positives & false negatives

To communicate effectively, **you have to “lie”!**
Expressing it in words: part 4

4. Sentence patterns & grammar. For instance, NLG system may need to choose between saying

1. *Sick toddlers and babies (...), or*
2. *Sick toddlers and sick babies (...)*

But (1) can also mean “Sick toddlers and all babies”, hence can misinform!

Which ambiguous sentences are likely to be misunderstood? How can misunderstandings be avoided? (Khan et al. 2012)
Aside: End-to-End systems that use Deep Learning systems (Dušek et al. 2018)

End-to-End = NLG without a pipeline

**Input:** name[Cotto], Type[coffee shop], near[The Bakers]

**Erroneous Text output** from 3 systems:

**GONG:** Cotto is a place near The Bakers. [“Omission”]

**SHEF2:** Cotto is a pub near The Bakers. [“Replacement”]

**TGEN:** Cotto is a coffee shop with a low price range. It is located near The Bakers. [“Hallucination”]

Deep Learning systems “lie” as well!
Three types of “lying”

1. Accidental lying

2. Benign deceit
   (on purpose, for the benefit of the reader only.
   “A white lie”. “Leugentje om bestwil”)

3. Taking advantage
   (on purpose, for the benefit of the speaker/system only.
   This is evil.)

So far: 1

Now briefly: 2
Benign deceit
Partial or incorrect info in the hearer’s interest

Babytalk BT-clan system:
- “Baby is OK”, to sick grandma (Moncur 2013)

NLG system should compare the likely effects of a truthful utterance versus a white lie (van Deemter & Reiter 2018)
- Use Crawford-Sobel Game Theory,
- Maximise expected utility
Maximising expected utility

This done by multiplying $a \times b$:

a. The probability of each possible effect of each possible utterance
   - e.g. the possible effects of saying “The baby is OK”, and the possible effects of saying “The baby is very unwell”

b. The utility of each of these effects

The idea is to generate that utterance where $a \times b$ is maximal.
Conclusion in 3 short parts:
Conclusion (1)

*NLG systems*

An emerging class of AI systems

– likely to have a profound impact on society
– medical, weather, journalism, economics, ...

But NLG systems also

– produce accidental falsehoods
– sometimes produce falsehoods on purpose
Conclusion (2)
Ian McEwan’s “Machines Like Me”

Deviating from the truth is not hard:
NLG (and human speakers?) do it all the time:

- Accidentally omitting or exaggerating trends
- Making imperfect simplifications
- Suggesting causal links that are doubtful, etc.

What is hard is speaking truthfully
Conclusion (3)

Disinformation / fake news

• Obama’s birthplace may be checked automatically

• But many “facts” are debatable
even when complete & correct data is available,
(as our discussion of the Babytalk system showed)

• Even worse when data are sparse:
  “CO₂ emissions are not causing global warming”

• How could such falsehoods be detected?

• **Deliberate** creation of false information is even harder to prove
Epilogue

Are these conclusions relativistic?

Yes: absolute truthfulness is often unachievable (e.g., like absolute chemical purity)

No: the ideal of truthfulness is essential to NLG systems, or else
  – patients suffer
  – citizens are misled, etc.
Epilogue

In journalism, politics, etc., the ideal of truthfulness is under treat (e.g. Bolsover & Howard 2018)

– Seen as unachievable and naive
– No longer worth striving for (?)

By contrast, Academia is a utopian society, devoted to the pursuit of knowledge, insight, and truth -- Difficult though that often is.
I am grateful to be a part of that society, with colleagues devoted to the same pursuit

[ Photos of the following people, shown from left to right and from top to bottom, were removed to save space ]


