Control in Automatic Text Summarization

Azamat Omuraliev
Research Thesis Intern @ NLING, ING
MSc Artificial Intelligence, University of Amsterdam

House rules:

- Please mute your mic if not speaking
- Ask your questions as messages
- Don’t hesitate to ask for clarifications!

ING

Plan for the presentation / Concepts covered

- Types of summarization
- Abstractive summarization
- Current issues
- Thesis research
  (Controllable Text Summarization)
- Summarization in industry
- QA

- *Seq2seq models (RNN, Word embeddings)*
- *Exposure bias*
- *Teacher forcing*
- *Optimizing for non-differentiable metrics*

- Feel free to ask questions throughout the presentation!

Example: where is ground truth and where is prediction?

Google Wallet says it has changed its policy when storing users' funds as they will now be federally-insured (file photo)

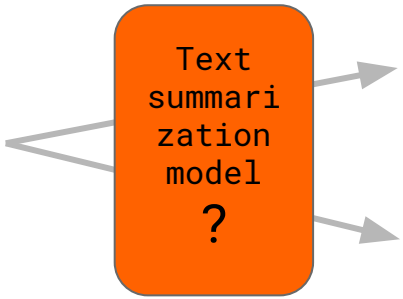
For those who use Google Wallet, their money just became safer with federal-level insurance.

Google confirmed to Yahoo Finance in a statement that its current policy changed - meaning the company will store the balances for users of the mobile transfer service (similar to PayPal and Venmo) in multiple federally-insured banking institutions.

This is good news for people who place large amounts of money in their Wallet Balance because the Federal Deposit Insurance Corporation insures funds for banking institutions up to $250,000.

Currently, Google's user agreement says funds are not protected by the FDIC.

However, a Google spokesperson told Yahoo Finance that the current policy has changed. (...)



Google spokesperson confirmed current policy changed meaning funds will be protected by the federal deposit insurance corporation. As a non-banking institution, Google Wallet, along with competitors PayPal and Venmo, is not legally required to be federally insured. With the new change to its policy, funds in wallet balance are protected if anything were to happen to the company like bankruptcy.

Google confirmed to Yahoo Finance in a statement that its current policy changed. The company will store the balances for users of the mobile transfer service (similar to PayPal and Venmo) in multiple federally-insured banking institutions. Google's user agreement says funds are not protected by the federal deposit insurance corporation.

Example: where is ground truth and where is prediction?

Google Wallet says it has changed its policy when storing users' funds as they will now be federally-insured (file photo)

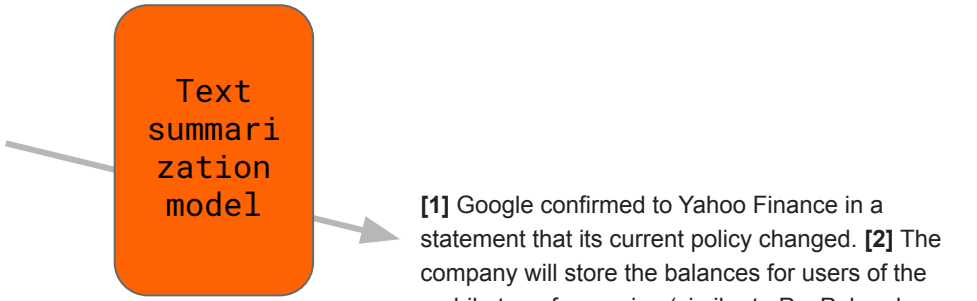
For those who use Google Wallet, their money just became safer with federal-level insurance.

[1] Google confirmed to Yahoo Finance in a statement that its current policy changed - meaning **[2] the company will store the balances for users of the mobile transfer service (similar to PayPal and Venmo) in multiple federally-insured banking institutions.**

This is good news for people who place large amounts of money in their Wallet Balance because the Federal Deposit Insurance Corporation insures funds for banking institutions up to $250,000.

Currently, **[3] Google's user agreement says funds are not protected by the FDIC**.

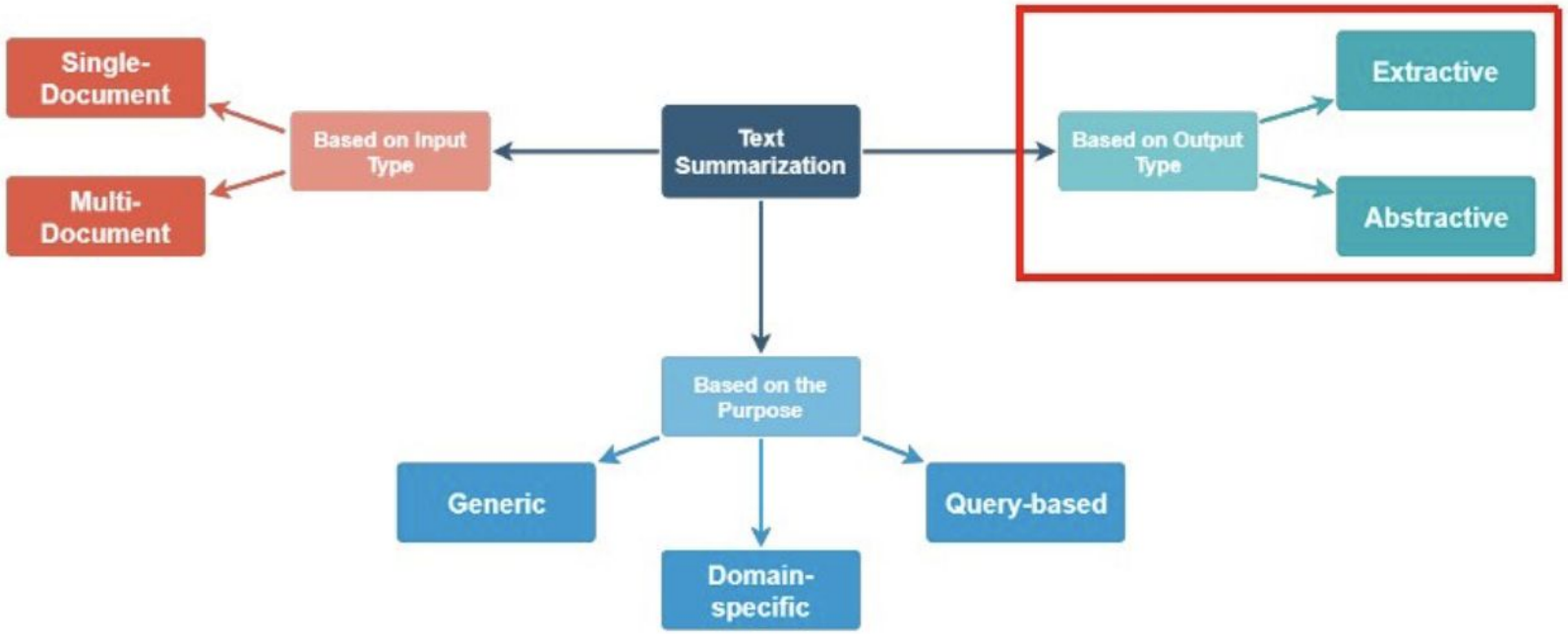
However, a Google spokesperson told Yahoo Finance that the current policy has changed. (...)



[1] Google confirmed to Yahoo Finance in a statement that its current policy changed. **[2]** The company will store the balances for users of the mobile transfer service (similar to PayPal and Venmo) in multiple federally-insured banking institutions. **[3]** Google's user agreement says funds are not protected by the federal deposit insurance corporation.

ING

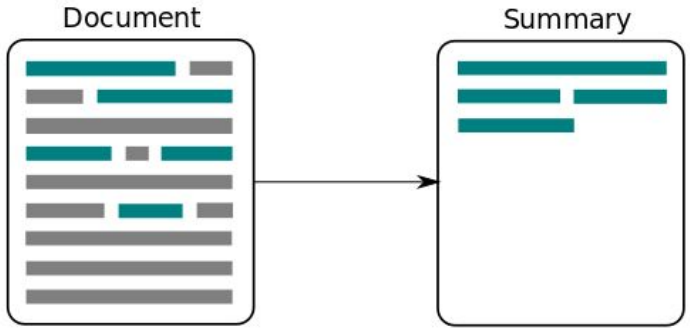
Types of text summarization



Process of summarizing text

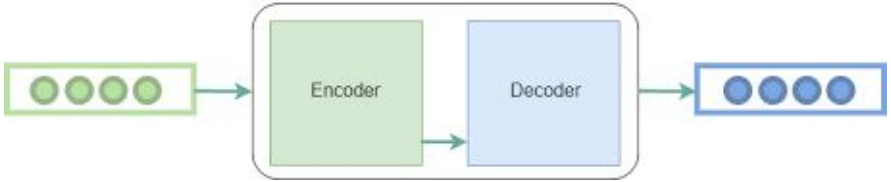
Extract

- Select a few sentences from the original document
- Make sure the selected sentences and the original document semantically



Abstract

- Predict the summary word by word, sequentially
- *Much harder!*



Predict sequences of words as summaries

Process of summarizing text

What we have:

- Text
- Documents
- Summaries

What do we need?

What we want:

- Model that creates summaries given text

Transform words to a numerical representation

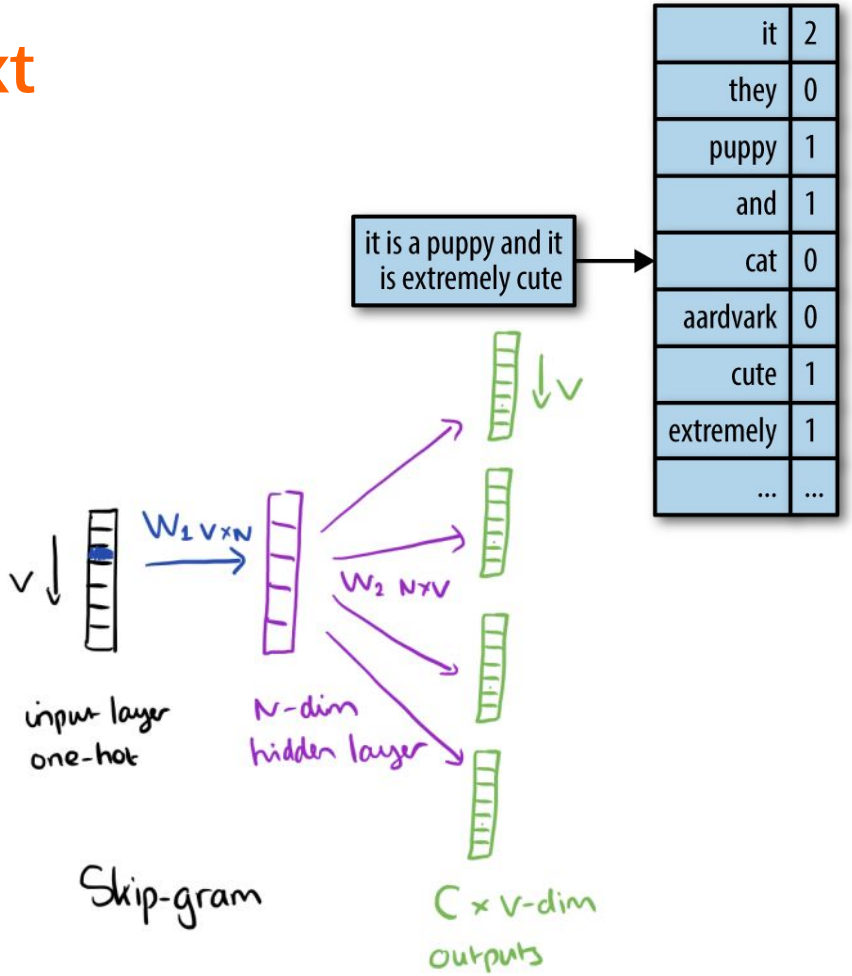
Combine words in the document to a single representation

Predict sequences of words as summaries

Process of summarizing text

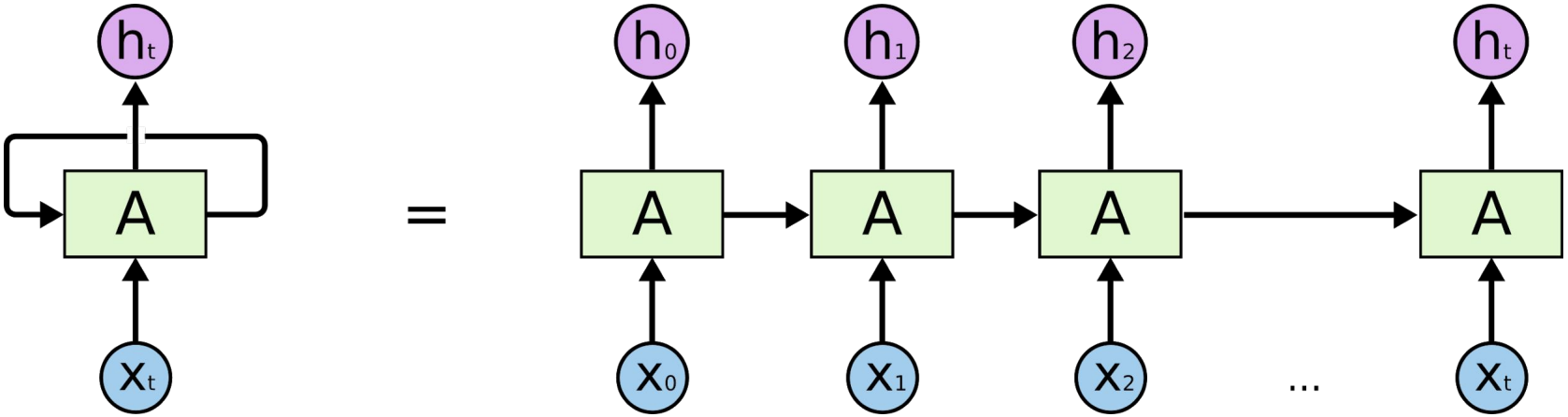
Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoret dolore magna aliquam erat volutpat. Ut

≠



Transform words to a numerical representation

Process of summarizing text



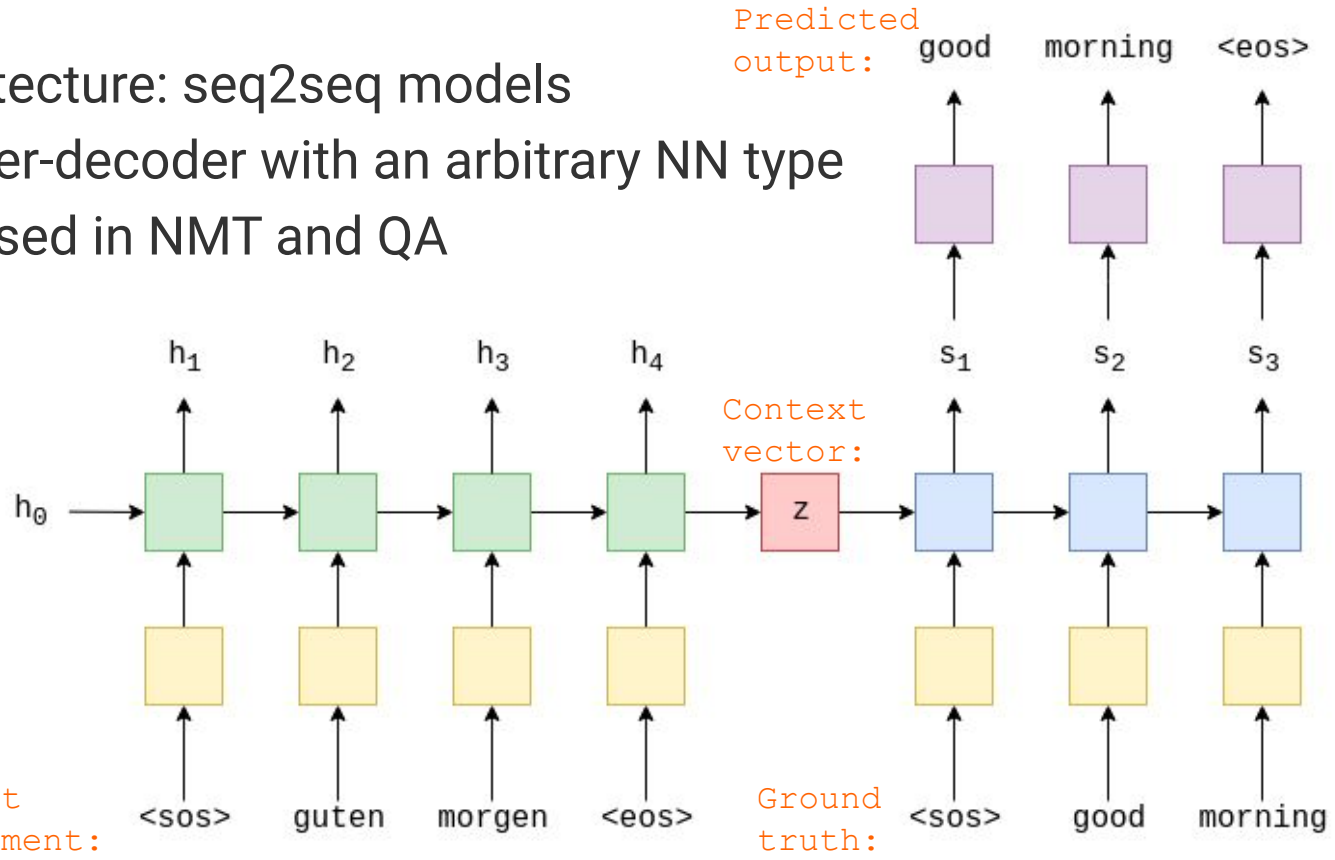
Combine words in the document to a single representation

Methods for text summarization: **Abstractive**

- Core architecture: seq2seq models
  - Encoder-decoder with an arbitrary NN type
  - Also used in NMT and QA

Green blocks = encoder
Blue blocks = decoder

Can be RNN, CNN, Transformer



https://github.com/bentrevett/pytorch-seq2seq

ING

Potential issues with abstractive summarization

Exposure bias

Output can be factually incorrect

?

Many potentially correct outputs

Potential issues with abstractive summarization

Exposure bias

Optimize directly for ROUGE

!

Output can be factually incorrect

Evaluate factual correctness

Many potentially correct outputs

Beam search

Tackling exposure bias with reinforcement learning

- Reinforcement learning can solve 2 issues in summarization:
  - Maximum likelihood ≠ Good ROUGE scores
  - Exposure bias
- Can we optimize directly for ROUGE?

$$\frac{number\_of\_overlapping\_words}{total\_words\_in\_reference\_summary}$$

- Yes! Use REINFORCE algorithm (*r* = ROUGE)

$$L_{rl} = (r(\hat{y}) - r(y^s)) \sum_{t=1}^{n'} \log p(y_t^s | y_1^s, \ldots, y_{t-1}^s, x)$$

Paulus, R., Xiong, C., & Socher, R. (2017). A deep reinforced model for abstractive summarization. *arXiv preprint arXiv:1705.04304*.

Current issues with text summarization

- Restricted to domains (e.g. news)
  - In terms of style, semantics, etc
- Only takes into account input text
  - No preferences, topics to look for, entities of focus
- Requires large amounts of data
- Not guaranteed to be factually correct
- Difficult to automatically evaluate (e.g. accuracy)
- Suffers from exposure bias

Controllable abstractive summarization

- Restricted to domains (e.g. news)
  - In terms of style, semantics, etc
- Only takes into account input text
  - No preferences, topics to look for, entities of focus

- Requires large amounts of data
- Not guaranteed to be factually correct
- Difficult to automatically evaluate (e.g. accuracy)
- Suffers from exposure bias

a. Summary with Length Control

Requesting Length 2: @entity0 [Easter] is over for the wild rabbits of greater @entity2 [Sydney] as councils and parks prepare another attempt to kill them off with a deadly virus. It comes after over 30 government bodies scattered carrots laced with calicivirus.

Requesting Length 6: @entity0 [Easter] is over for the wild rabbits of greater @entity2 [Sydney] as councils and parks prepare another attempt to kill them off with a deadly virus. This year, because of really high summer rainfall - which led to great food availability - there has been a big surge in the rabbit population in @entity2 [Sydney].

Requesting Length 10: @entity0 [Easter] is over for the wild rabbits of greater @entity2 [Sydney] as councils and parks prepare another attempt to kill them off with strategically placed carrots that have been laced with a deadly virus. This year,because of really high summer rainfall - which led to great food availability - there has been a big surge in the rabbit population in @entity2 [Sydney]. It comes after over 30 government bodies scattered carrots laced with calicivirus around public areas in March.

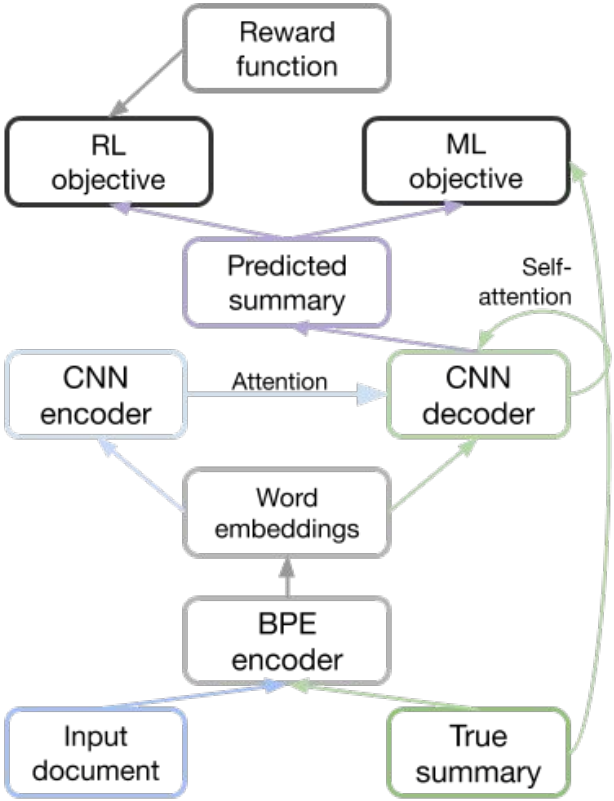
Thesis research: Control in text summarization

- Optimize for maximum likelihood
- Optimize for quantifiable control function, used as reward

$$L_{rl} = (r(\hat{w}) - r(w^s)) \log p_\theta(w^s) =$$

$$= (r(\hat{w}) - r(w^s)) \sum_{t=1}^{T} \log p_\theta(w_t^s | w_1^s, ..., w_{t-1}^s, x)$$

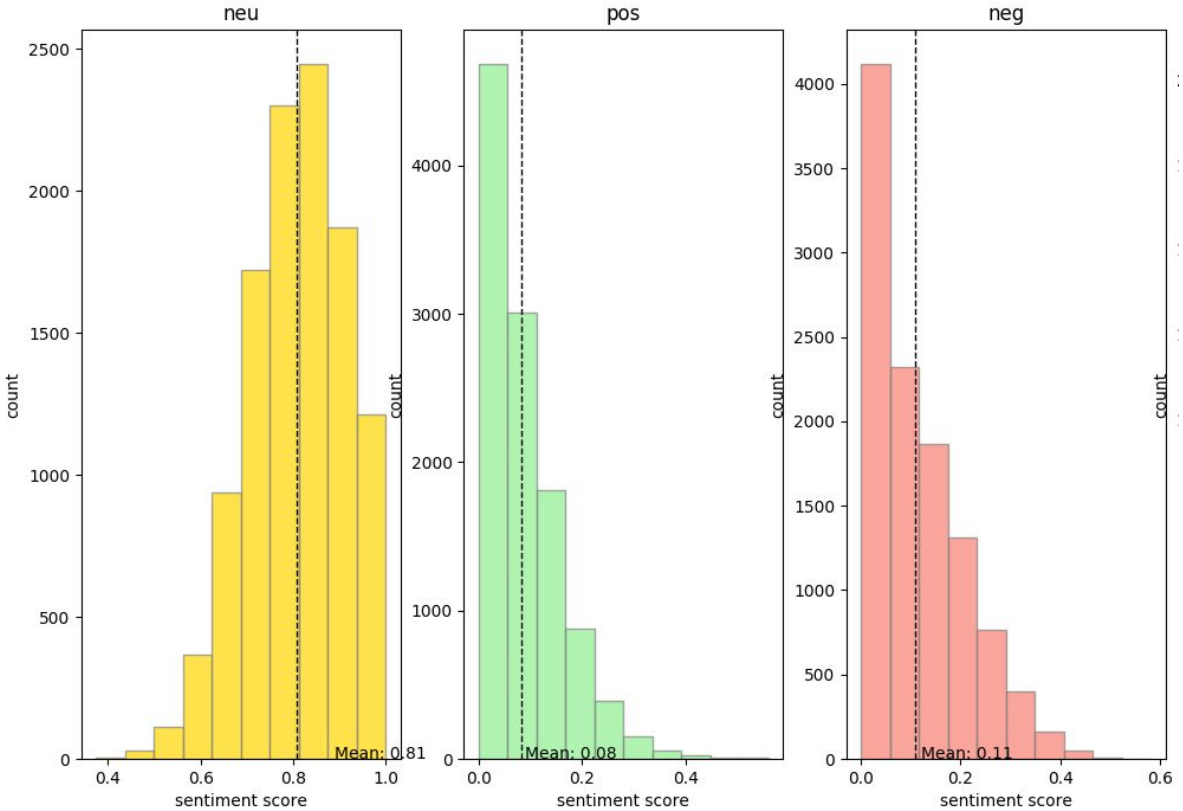
- In past, *r* = ROUGE
- Contribution: *r* as sentiment, vocabulary count, length of words or sentences



ING

Control on out-of-domain features

Sentiment in summaries from CNN/Daily Mail dataset



Thesis research: Control in text summarization

- Reinforcement learning poses fewer requirements on reward function
  - *r* should not be constant
  - *r* should change when policy θ changes
- Previously: control as a feature in conditional learning
  - Now: control as an out-of-domain feature
- Link to controlled natural language

What to expect?

- Replace valent words with nonvalent synonyms (*reward*)
- Maintain natural language coherence (*maximum likelihood*)

Words	Agg. scores	Ind. scores	Freq.	Synonyms	Syn. scores
help	631.161845	1.7	395	['aid', 'assist', 'assistance']	[0, 0, 0]
like	435.09055	1.5	337	['the_like', 'the_likes_of']	[0, 0]
win	787.668	2.8	289	['winnings', 'profits']	[2.5, 1.9]
top	195.024	0.8	253	['top_side', 'upper_side', 'upside']	[0, 0, 0]
won	514.08435	2.7	204	['South_Korean_won']	[0]
security	284.31975	1.4	203	['protection']	[0]
support	295.2185	1.7	189	['reinforcement', 'reenforcement']	[0, 0]
want	38.5219	0.3	185	['privation', 'deprivation', 'neediness']	[0, -1.8, 0]
good	306.3397	1.9	179	['goodness']	[2.0]
wins	479.25	2.7	179	['win']	[2.8]

Applications in industry

- Salesforce [link]
  - Use summarization to recap emails and customer reviews
- Lionbridge AI [link]
  - Localize Japanese hotel page content as a summary in English
- Semantic MEDLINE
  - Secondary database curation in genetic research
- Ebrevia [link]
  - Summarize contracts for the review process in legal domain
- IBM and Mphasis provide summarization APIs

Workman, T. E., Fiszman, M., Hurdle, J. F., & Rindflesch, T. C. (2010). Biomedical text summarization to support genetic database curation: using Semantic MEDLINE to create a secondary database of genetic information. *Journal of the Medical Library Association: JMLA*, *98*(4), 273.

Thanks for your attention!
Questions?

ING