Natural Language Processing 1

Lecture 9: Language generation and summarisation

Katia Shutova

ILLC
University of Amsterdam

26 November 2020

Language generation

Text summarisation

Extractive summarisation

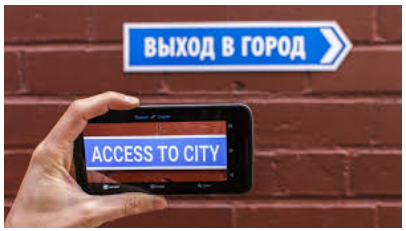
Query-focused multi-document summarisation

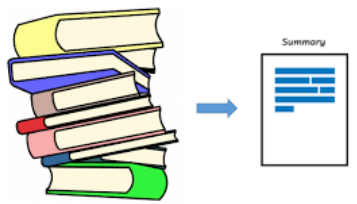
Summarisation using neural networks

Evaluating summarisation systems

Language generation tasks

- Dialogue modelling
- Email answering
- Machine translation
- Summarisation
- and many others





Language generation

Generation from what?! (Yorick Wilks)

Generation: some starting points

- Some semantic representation:
  - logical form (early work)
  - distributional representations (e.g. paraphrasing)
  - hidden states of a neural network
- Formally-defined data: databases, knowledge bases
- Numerical data: e.g., weather reports.

Regeneration: transforming text

- Machine translation
- Paraphrasing
- Summarisation
- Text simplification

Subtasks in generation

- Content selection: deciding what information to convey (selecting important or relevant content)
- Discourse structuring: overall ordering
- Aggregation: splitting information into sentence-sized chunks
- Referring expression generation: deciding when to use pronouns, which modifiers to use etc
- Lexical choice: which lexical items convey a given concept
- Realisation: mapping from a meaning representation to a string
- Fluency ranking: discriminate between grammatically / semantically valid and invalid sentences

Approaches to generation

- Templates: fixed text with slots, fixed rules for content selection.
- Statistical: use machine learning (supervised or unsupervised) for the various subtasks.
- Deep learning: particularly for regeneration tasks.

Large scale dialogue and question answering systems, such as Siri, use a combination of the above techniques.

Text summarisation

Task: generate a short version of a text that contains the most important information

Single-document summarisation:

- given a single document
- produce its short summary

Multi-document summarisation:

- given a set of documents
- produce a brief summary of their content

Generic vs. Query-focused summarisation

Generic summarisation:

- identifying important information in the document(s) and presenting it in a short summary

Query-focused summarisation:

- summarising the document in order to answer a specific query from a user

A simple example of query-focused summarisation

Google

what is natural language processing?

Web Videos News Images Shopping More Search tools

About 30,200,000 results (0.38 seconds)

Natural language processing (**NLP**) is a field of computer science, artificial intelligence, and computational linguistics concerned with the interactions between computers and human (**natural**) languages. As such, **NLP** is related to the area of human–computer interaction.

Natural language processing - Wikipedia, the free ...
https://en.wikipedia.org/wiki/**Natural_language_processing**

Feedback

Natural language processing - Wikipedia, the free ...
https://en.wikipedia.org/wiki/**Natural_language_processing**
Natural language processing (**NLP**) is a field of computer science, artificial intelligence, and computational linguistics concerned with the interactions between computers and human (**natural**) languages. As such, **NLP** is related to the area of human–computer interaction.
Outline of natural language ... - Natural language understanding

Approaches

Extractive summarisation:

- extract important / relevant sentences from the document(s)
- combine them into a summary

Abstractive summarisation:

- interpret the content of the document (semantics, discourse etc.) and generate the summary
- formulate the summary using other words than in the document
- very hard to do!

Extractive summarisation

decide what sentence should be in the summary, what should not be in the summary

Three main components:

- Content selection: identify important sentences to extract from the document
- Information ordering: order the sentences within the summary
- Sentence realisation: sentence simplification

Content selection – unsupervised approach

- Choose sentences that contain informative words
- Informativeness measured by:
  - tf-idf: assign a weight to each word i in the doc j as

term frequency. freq of a word in a doc

$$weight(w_i) = tf_{ij} * idf_i$$

tf_{ij} – frequency of word i in doc j
idf_i – inverse document frequency

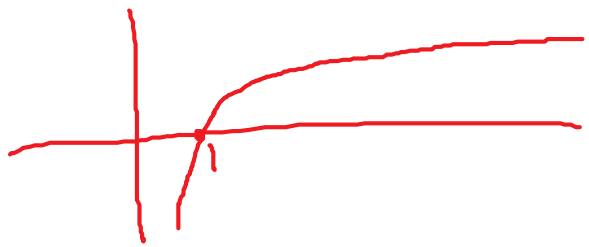
how many doc the word occurs

$$idf_i = \log \frac{N}{n_i}$$

total nb of doc in corpus

nb of doc that the word occur

N – total docs; n_i docs containing w_i



  - mutual information

Content selection – supervised approach

- start with a training set of documents and their summaries
- align sentences in summaries and documents
- extract features:
  - position of the sentence (e.g. first sentence)
  - sentence length
  - informative words
  - cue phrases
  - etc.
- train a binary classifier: should the sentence be included in the summary?

Content selection – supervised vs. unsupervised

Problems with the supervised approach:

- difficult to obtain data
- difficult to align human-produced summaries with sentences in the doc
- doesn't perform better than unsupervised in practice

Ordering sentences

For single-document summarisation:

- very straightforward
- simply follow the order in the original document

An example summary

from Nenkova and McKeown (2011):

As his lawyers in London tried to quash a Spanish arrest warrant for Gen. Augusto Pinochet, the former Chilean Dictator, efforts began in Geneva and Paris to have him extradited. Britain has defended its arrest of Gen. Augusto Pinochet, with one lawmaker saying that Chile's claim that the former Chilean Dictator has diplomatic immunity is ridiculous. Margaret Thatcher entertained former Chilean Dictator Gen. Augusto Pinochet at her home two weeks before he was arrested in his bed in a London hospital, the ex-prime minister's office said Tuesday, amid growing diplomatic and domestic controversy over the move.

Query-focused multi-document summarisation

Example query: "*Describe the coal mine accidents in China and actions taken*"

Steps in summarization:

1. find a set of relevant documents
2. simplify sentences
3. identify informative sentences in the documents
4. order the sentences into a summary
5. modify the sentences as needed

Sentence simplification

- parse sentences
- hand-code rules to decide which modifiers to prune
  - appositives: e.g. *Also on display was a painting by Sandor Landeau, ~~an artist who was living in Paris at the time.~~*
  - attribution clauses: e.g. *Eating too much bacon can lead to cancer, ~~the WHO reported on Monday.~~*
  - PPs without proper names: e.g. *Electoral support for Plaid Cymru increased ~~to a new level.~~*
  - initial adverbials: e.g. *~~For example,~~ ~~On the other hand,~~*
- also possible to develop a classifier (e.g. satelite identification and removal)

Content selection from multiple documents

Select informative and non-redundunt sentences:

- Estimate informativeness of each sentence (based on informative words)
- Start with the most informative sentence:
  - identify informative words based on e.g. tf-idf
  - words in the query also considered informative
- Add sentences to the summary based on maximal marginal relevance (MMR)

Content selection from multiple documents

Maximal marginal relevance (MMR): iterative method to choose the best sentence to add to the summary so far

- Relevance to the query: high cosine similarity between the sentence and the query
- Novelty wrt the summary so far: low cosine similarity with the summary sentences

$$\hat{s} = \underset{s_i \in D}{\text{argmax}} \left[\lambda sim(s_i, Q) - (1 - \lambda) \max_{s_j \in S} sim(s_i, s_j) \right]$$

Stop when the summary has reached the desired length

Sentence ordering in the summary

- Chronologically: e.g. by date of the document
- Coherence:
  - order based on sentence similarity (sentences next to each other should be similar, e.g. by cosine)
  - order so that the sentences next to each other discuss the same entity / referent
- Topical ordering: learn a set of topics present in the documents, e.g. using topic modelling, and then order sentences by topic.

Example summary

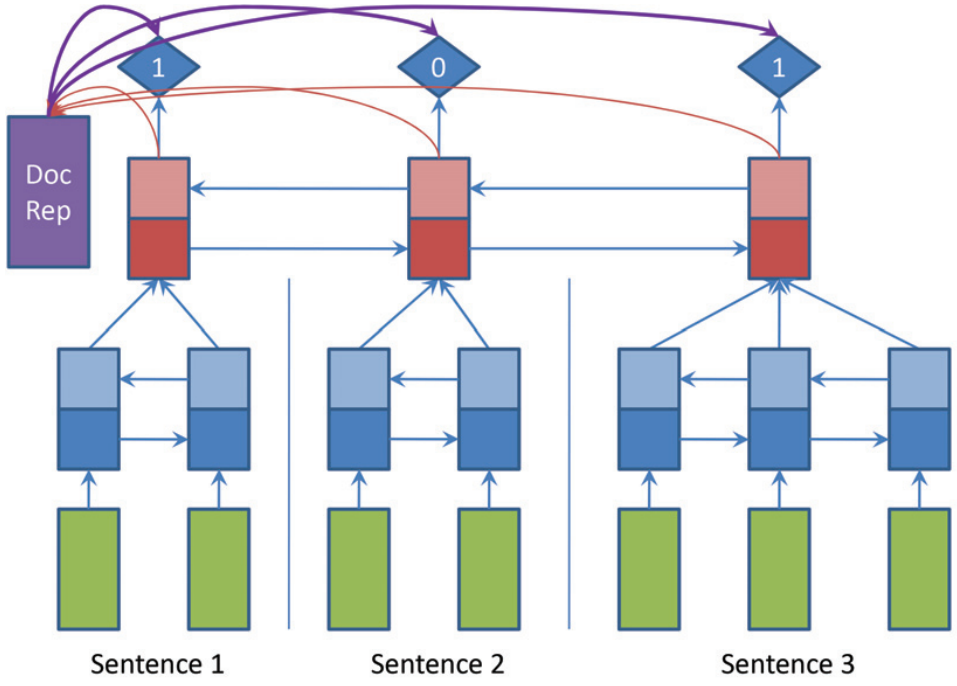
Query: "*Describe the coal mine accidents in China and actions taken*"

Example summary (from Li and Li 2013):
(1) In the first eight months, the death toll of coal mine accidents across China rose 8.5 percent from the same period last year. (2) China will close down a number of ill-operated coal mines at the end of this month, said a work safety official here Monday. (3) Li Yizhong, director of the National Bureau of Production Safety Supervision and Administration, has said the collusion between mine owners and officials is to be condemned. (4) from January to September this year, 4,228 people were killed in 2,337 coal mine accidents. (5) Chen said officials who refused to register their stakes in coal mines within the required time

Extractive summarisation with RNNs

Nallapati et al. 2017. *SummaRuNNer: A Recurrent Neural Network Based Sequence Model for Extractive Summarization of Documents*

- Use an RNN to build a representation of a document
- Classify sentences in the document as 0 or 1 (included in the summary or not)



SummaRuNNer

Document representation:

$$\mathbf{d} = \tanh(W_d \frac{1}{N_d} \sum_{j=1}^{N^d} [\mathbf{h}_j^f, \mathbf{h}_j^b] + \mathbf{b}),$$

Computing the label probability for a sentence:

$$\begin{aligned} P(y_j = 1|\mathbf{h}_j, \mathbf{s}_j, \mathbf{d}) = \sigma(&W_c \mathbf{h}_j && \#\texttt{(content)} \\ &+ \mathbf{h}_j^T W_s \mathbf{d} && \#\texttt{(salience)} \\ &- \mathbf{h}_j^T W_r \tanh(\mathbf{s_j}) && \#\texttt{(novelty)} \\ &+ b), && \#\texttt{(bias term)} \end{aligned}$$

Representation of the summary so far:

$$\mathbf{s}_j = \sum_{i=1}^{j-1} \mathbf{h}_i P(y_i = 1|\mathbf{h}_i, \mathbf{s}_i, \mathbf{d}).$$

SummaRuNNer

Document representation:

$$\mathbf{d} = \tanh(W_d \frac{1}{N_d} \sum_{j=1}^{N^d} [\mathbf{h}_j^f, \mathbf{h}_j^b] + \mathbf{b}),$$

Computing the label probability for a sentence:

$$\begin{aligned} P(y_j = 1 | \mathbf{h}_j, \mathbf{s}_j, \mathbf{d}) = \sigma(&W_c \mathbf{h}_j && \#\texttt{(content)} \\ &+ \mathbf{h}_j^T W_s \mathbf{d} && \#\texttt{(salience)} \\ &- \mathbf{h}_j^T W_r \tanh(\mathbf{s_j}) && \#\texttt{(novelty)} \\ &+ b), && \#\texttt{(bias term)} \end{aligned}$$

Representation of the summary so far:

$$\mathbf{s}_j = \sum_{i=1}^{j-1} \mathbf{h}_i P(y_i = 1 | \mathbf{h}_i, \mathbf{s}_i, \mathbf{d}).$$

SummaRuNNer

Document representation:

$$\mathbf{d} = \tanh(W_d \frac{1}{N_d} \sum_{j=1}^{N^d} [\mathbf{h}_j^f, \mathbf{h}_j^b] + \mathbf{b}),$$

Computing the label probability for a sentence:

$$\begin{aligned} P(y_j = 1 | \mathbf{h}_j, \mathbf{s}_j, \mathbf{d}) = \sigma(&W_c \mathbf{h}_j && \#\texttt{(content)} \\ &+ \mathbf{h}_j^T W_s \mathbf{d} && \#\texttt{(salience)} \\ &- \mathbf{h}_j^T W_r \tanh(\mathbf{s_j}) && \#\texttt{(novelty)} \\ &+ b), && \#\texttt{(bias term)} \end{aligned}$$

Representation of the summary so far:

$$\mathbf{s}_j = \sum_{i=1}^{j-1} \mathbf{h}_i P(y_i = 1 | \mathbf{h}_i, \mathbf{s}_i, \mathbf{d}).$$

Abstractive summarisation

Task: given a short article, generate a **headline**

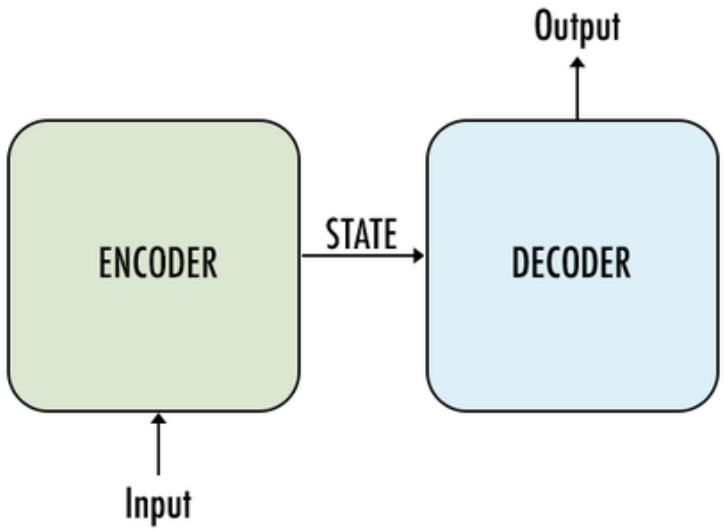
Training data: e.g. Gigaword (10m articles), CNN dataset

Article	Human Generated Headline
Usain Bolt rounded off the world championships Sunday by claiming his third gold in Moscow as he anchored Jamaica to victory in the men's 4x100m relay. The fastest man in the world charged clear of United States rival Justin Gatlin as the Jamaican quartet of Nesta Carter, Kemar Bailey-Cole, Nickel Ashmeade and Bolt won in 37.36 seconds.	Usain Bolt wins third gold of world championship
A ferocious leopard may have killed 15 people in Nepal in a 15-month span, its latest victim a 4-year-old boy that the creature dragged away into the jungle to eat. The head of boy was found in the forest a kilometer from his home Saturday morning, said Kamal Prasad Kharel, the police chief of the Baitadi district, an area about 600 kilometers (373 miles) west of Kathmandu.	A 4-year-old boy is the latest victim of a man-eating leopard, a local police chief says

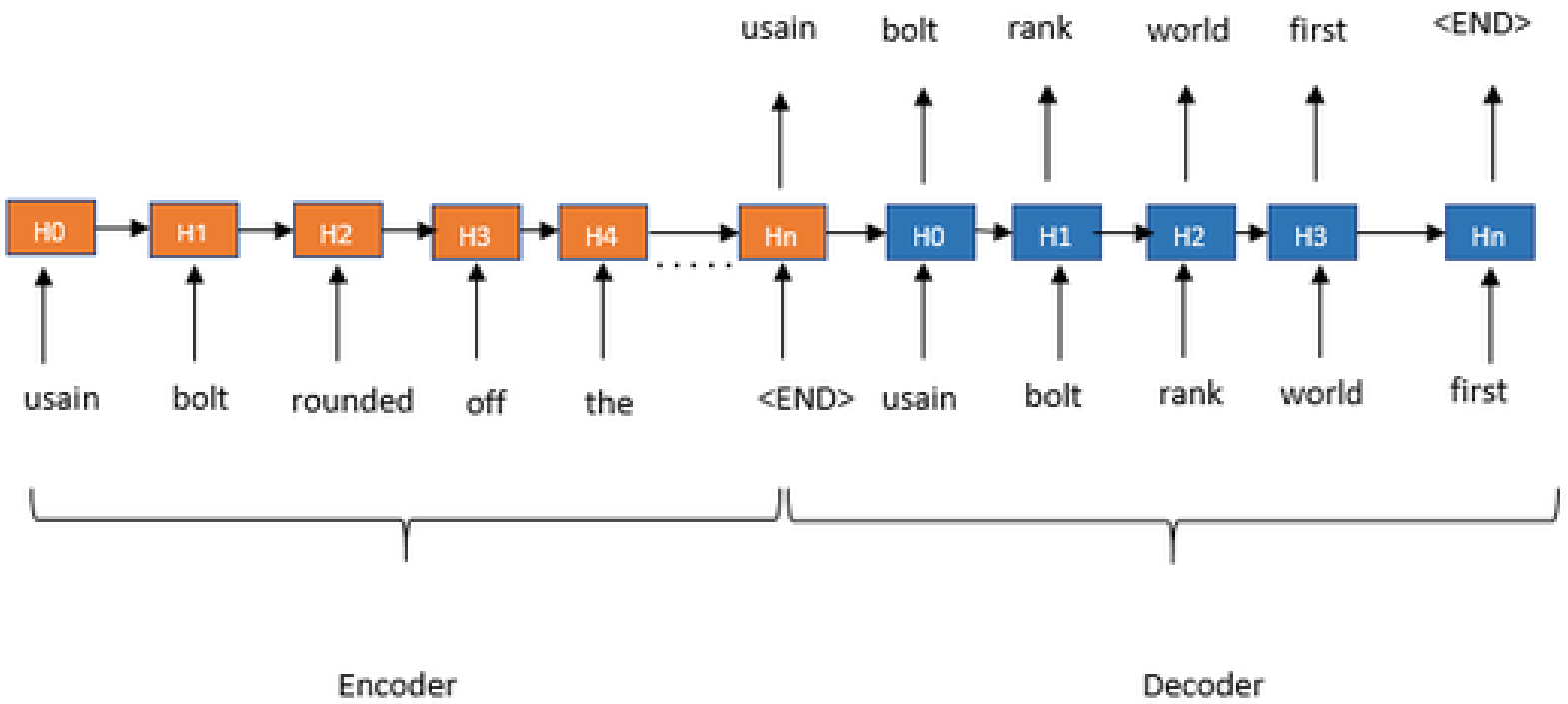
Abstractive summarisation with RNNs

Sequence-to-sequence models:

- **Encoder RNN**: produces a fixed-size vector representation of the input document
- **Decoder RNN**: generates the output summary word-by-word based on the input representation



Sequence-to-sequence models



Example summaries

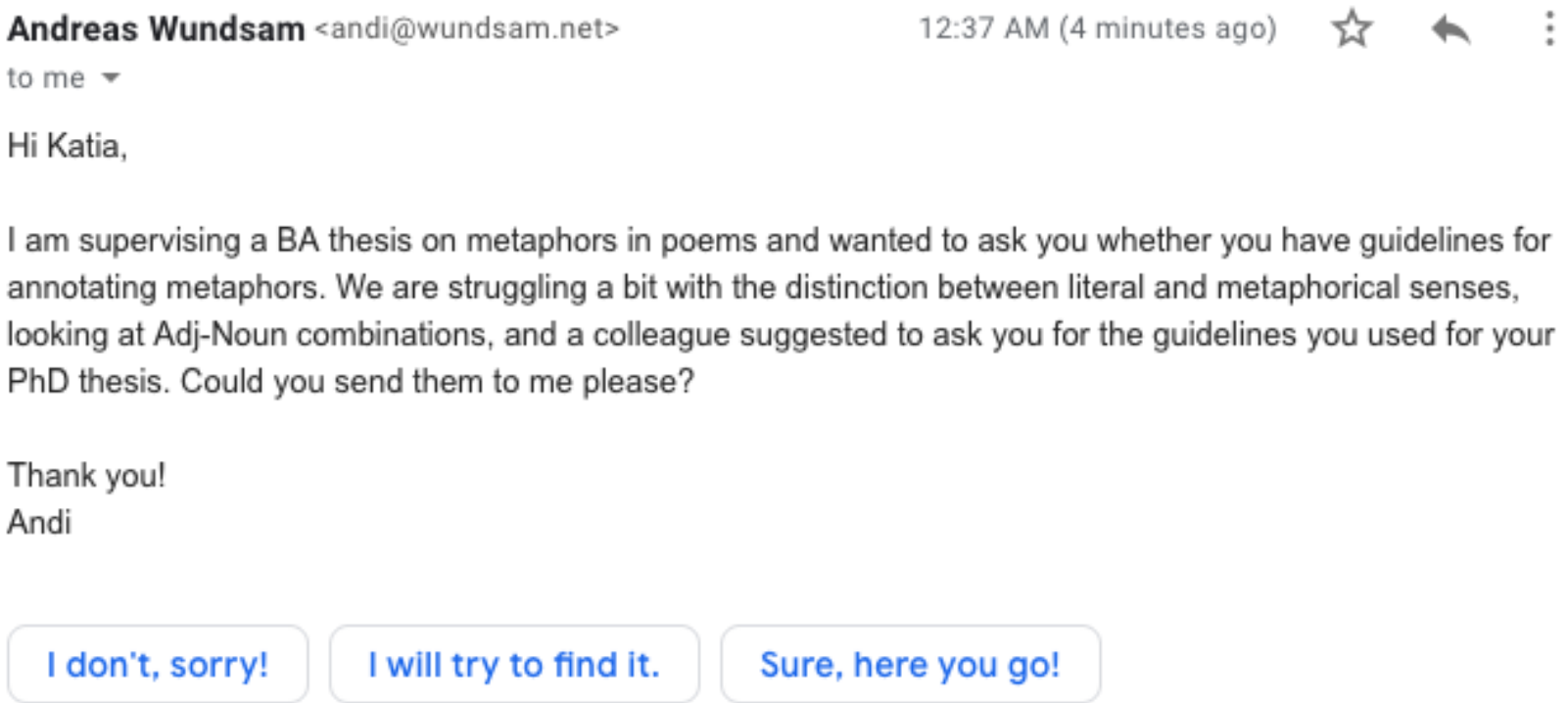
Chopra et al. 2017. *Abstractive Sentence Summarization with Attentive Recurrent Neural Networks*

Input: economic growth in toronto will suffer this year because of sars, a think tank said friday as health authorities insisted the illness was under control in canada's largest city.
Summary: *think tank says economic growth in toronto will suffer this year*

Input: an international terror suspect who had been under a controversial loose form of house arrest is on the run, british home secretary john reid said tuesday.
Summary: *international terror suspect under house arrest*

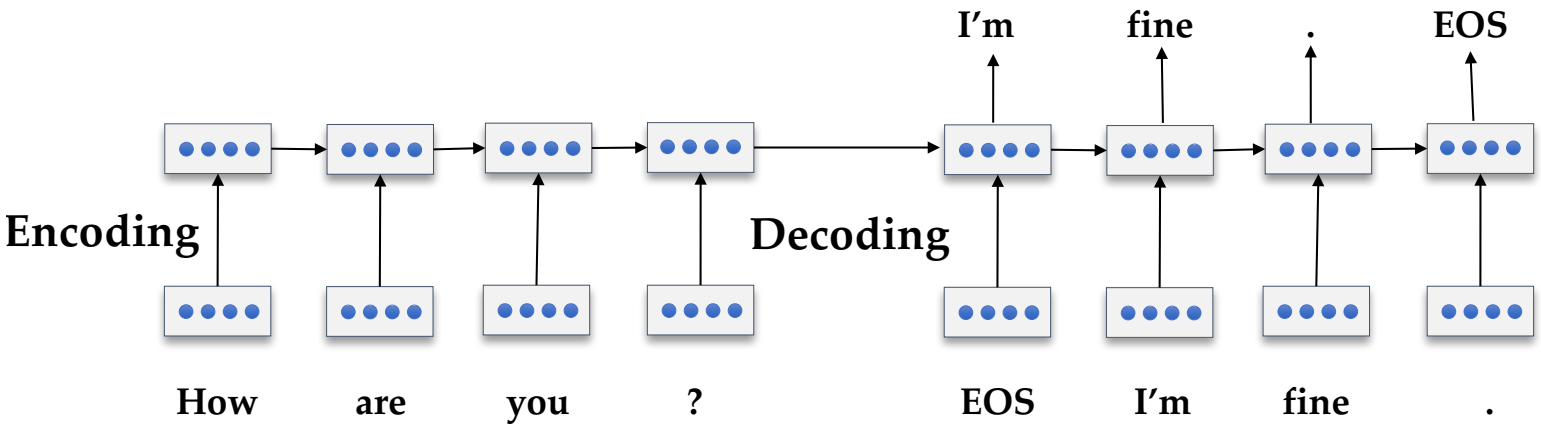
Other applications of *seq2seq* models

Email answering: Google's *Smart Reply* feature



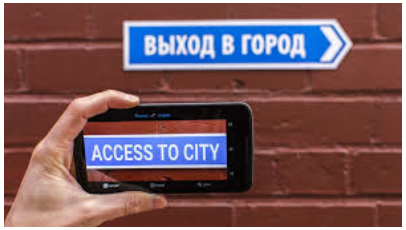
Other applications of *seq2seq* models

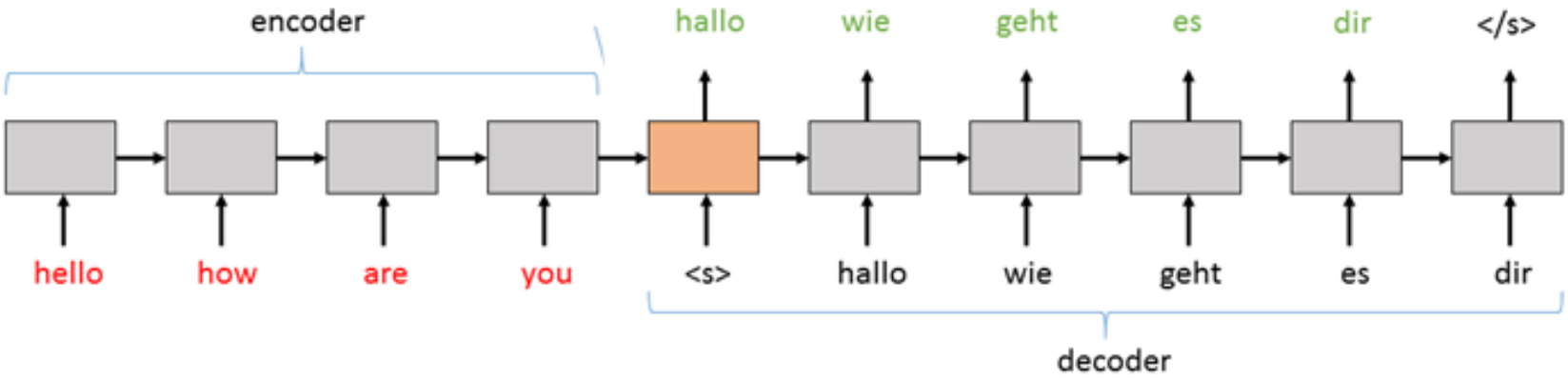
Dialogue modelling



Other applications of *seq2seq* models

Machine translation





Evaluating summarisation systems

1. Evaluate against human judgements
   - "Is this a good summary?"
   - Use multiple subjects, measure agreement
   - The best way, but expensive

2. ROUGE (Recall oriented understudy for gisting evaluation)
   For each document in the dataset:
   - humans produce a set of reference summaries $R_1, ..., R_N$
   - the system generates a summary S
   - compute the percentage of n-grams from the reference summaries that occur in S

ROUGE

- let's look at ROUGE-2 — using bigrams
- compute the percentage of bigrams from the reference summaries $R_1, ..., R_N$ that occur in S

$$\text{ROUGE-2} = \frac{\sum_{R_i} \sum_{bigram_j \in R_i} count_{match}(j, S)}{\sum_{R_i} \sum_{bigram_j \in R_i} count(j, R_i)}$$

ROUGE example

Question: "What is dadaism?"

Human 1: Dadaism was an art movement formed during the First World War in Zurich in negative reaction to the horrors of the war.

Human 2: Dada or Dadaism was a form of artistic anarchy born out of disgust for the social, political and cultural values of the time.

Human 3: Dadaism was a short-lived but highly influential art movement from the early 20th century.

System: Dada or Dadaism was an art movement of the European avant-garde in the early 20th century.

$$\text{ROUGE-2} = \frac{}{21 + 22 + 13}$$

ROUGE example

Question: "What is dadaism?"

Human 1: Dadaism was an art movement formed during the First World War in Zurich in negative reaction to the horrors of the war.

Human 2: Dada or Dadaism was a form of artistic anarchy born out of disgust for the social, political and cultural values of the time.

Human 3: Dadaism was a short-lived but highly influential art movement from the early 20th century.

System: Dada or Dadaism was an art movement of the European avant-garde in the early 20th century.

$$\text{ROUGE-2} = \frac{}{21 + 22 + 13}$$

ROUGE example

Question: "What is dadaism?"

Human 1: **Dadaism was an art movement** formed during the First World War in Zurich in negative reaction to the horrors **of the** war.

Human 2: Dada or Dadaism was a form of artistic anarchy born out of disgust for the social, political and cultural values of the time.

Human 3: Dadaism was a short-lived but highly influential art movement from the early 20th century.

System: Dada or **Dadaism was an art movement of the** European avant-garde in the early 20th century.

$$\text{ROUGE-2} = \frac{5+}{21+22+13}$$

ROUGE example

Question: "What is dadaism?"

Human 1: Dadaism was an art movement formed during the First World War in Zurich in negative reaction to the horrors of the war.

Human 2: **Dada or Dadaism was** a form of artistic anarchy born out of disgust for the social, political and cultural values **of the** time.

Human 3: Dadaism was a short-lived but highly influential art movement from the early 20th century.

System: **Dada or Dadaism was** an art movement **of the** European avant-garde in the early 20th century.

$$\text{ROUGE-2} = \frac{5 + 4 +}{21 + 22 + 13}$$

ROUGE example

Question: "What is dadaism?"

Human 1: Dadaism was an art movement formed during the First World War in Zurich in negative reaction to the horrors of the war.

Human 2: Dada or Dadaism was a form of artistic anarchy born out of disgust for the social, political and cultural values of the time.

Human 3: **Dadaism was** a short-lived but highly influential **art movement** from **the early 20th century**.

System: Dada or **Dadaism was** an **art movement** of the European avant-garde in **the early 20th century**.

$$\text{ROUGE-2} = \frac{5 + 4 + 5}{21 + 22 + 13}$$

ROUGE example

Question: "What is dadaism?"

Human 1: Dadaism was an art movement formed during the First World War in Zurich in negative reaction to the horrors of the war.

Human 2: Dada or Dadaism was a form of artistic anarchy born out of disgust for the social, political and cultural values of the time.

Human 3: **Dadaism was** a short-lived but highly influential **art movement** from **the early 20th century**.

System: Dada or **Dadaism was** an **art movement** of the European avant-garde in **the early 20th century**.

$$\text{ROUGE-2} = \frac{5+4+5}{21+22+13} = \frac{14}{56} = 0.25$$

State of the art in summarisation

Dong, 2018. *A Survey on Neural Network-Based Summarization Methods*

- Extractive summarisation

$$\text{The highest ROUGE-2} = 0.27$$

- Abstractive summarisation

$$\text{The highest ROUGE-2} = 0.17$$

Though the task / datasets are different, so not directly comparable.

Acknowledgement

Some slides were adapted from Dan Jurafsky