

# Lecture 10: Language generation

Overview of Natural Language Generation

Text summarisation

Extractive summarisation

Query-focused multi-document summarisation

## Overview of Natural Language Generation

Text summarisation

Extractive summarisation

Query-focused multi-document summarisation

# Language generation

Generation from what?! (Yorick Wilks)

# Generation

## Starting points:

- ▶ Some semantic representation:
  - ▶ logical form (early work)
  - ▶ distributional representations (e.g. paraphrasing)
- ▶ Formally-defined data: databases, knowledge bases
- ▶ Semi-structured data: tables, graphs etc.
- ▶ Numerical data: e.g., weather reports.
- ▶ User input in assistive communication.

## Regeneration: transforming text

- ▶ Statistical machine translation
- ▶ Paraphrasing
- ▶ Summarization
- ▶ Text simplification

## Components of a generation system

**Content selection** deciding what information to convey  
(selecting important or relevant content)

**Discourse structuring** overall ordering, sub-headings etc

**Aggregation** deciding how to split 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

**Realization** mapping from a meaning representation (or syntax  
tree) to a string (or speech)

**Fluency ranking** discriminate between grammatically /  
semantically valid and invalid sentences

## Approaches to generation

- ▶ Early work (limited domain): **hand-written rules** for first five steps, grammar for realization
- ▶ **Templates** (limited domain): most practical systems. Fixed text with slots, fixed rules for content selection.
- ▶ **Statistical** (limited domain): components as above, but use machine learning (supervised or unsupervised).
- ▶ **Regeneration**: statistical or mixed.

## Overview of Natural Language Generation

### Text summarisation

#### Extractive summarisation

#### Query-focused multi-document summarisation



# Text summarisation <sup>1</sup>

**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

---

<sup>1</sup>This part of the lecture is based on Dan Jurafsky's summarisation lecture, and is (quite appropriately) a summary thereof. The full lecture can be viewed online at <https://class.coursera.org/nlp/lecture/preview>

## 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)  
[https://en.wikipedia.org/wiki/Natural\\_language\\_processing](https://en.wikipedia.org/wiki/Natural_language_processing)

*Feedback*

## Natural language processing - Wikipedia, the free ...

[https://en.wikipedia.org/wiki/Natural\\_language\\_processing](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!

## Overview of Natural Language Generation

Text summarisation

**Extractive summarisation**

Query-focused multi-document summarisation

## Extractive summarisation

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

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

$tf_{ij}$  – frequency of word  $i$  in doc  $j$

$idf_i$  – inverse document frequency

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

$N$  – total docs;  $n_i$  docs containing  $w_i$

- ▶ **mutual information**
- ▶ **log-likelihood ratio** (LLR)

## 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

## 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.*

## Overview of Natural Language Generation

Text summarisation

Extractive summarisation

Query-focused multi-document summarisation

## 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.~~*
  - ▶ **PBs 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. satellite identification and removal)

## Content selection from multiple documents

Select **informative** and **non-redundant** 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

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

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 LDA, 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*