# Natural Language Processing

Dr. Chetana Gavankar, Ph.D,
IIT Bombay Monash University Australia
Chetana.gavankar@pilani.bitstipilani.ac.in

BITS Pilani
Pilani Campus

BITS Pilani
Pilani Campus

# Session 15 Text Summarization
# Date – 23rd March 2024

These slides are prepared by the instructor, with grateful acknowledgement of Prof. Dan Jurafsky and many others who made their course materials freely available online.

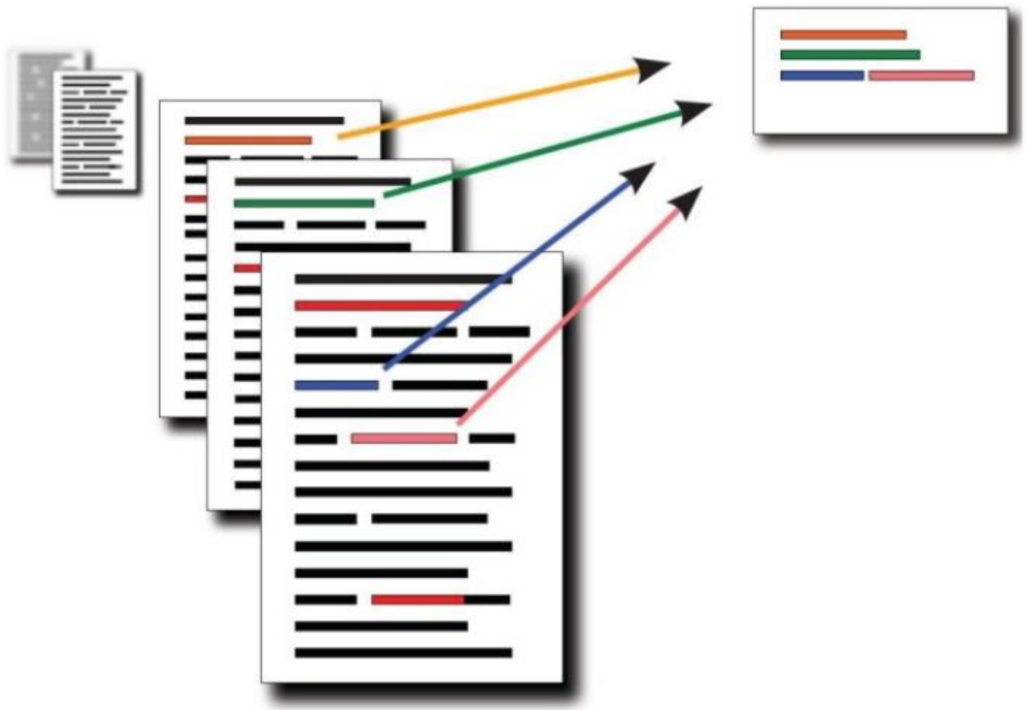
# Session Content

- What is Text Summarization?
- Applications
- Type of Summarization
  - Single Document Summarization
  - Multidocument Summarization
  - Extractive Summarization
  - Abstractive Summarization
  - Generic Summarization
  - Query focused summarization
- Stages of Summarization
  - Content Selection
  - Information Ordering
  - Sentence Realization

•Neural Text Summarization

# What is Text Summarization?

Task: produce an abridged version of a text while retaining the key, relevant information



# Applications

Useful for creating

- outlines or abstracts of any document, article, etc
- summaries of chat and email
- action items from a meeting
- simplifying text by compressing sentences

# Text Summarization

Input:

- single document summarization (SDS)
- multiple-document summarization (MDS)

Output:

- extractive
- abstractive

Focus:

- generic (unconditioned)
- query-focused (conditioned)

Approach:

- supervised
- unsupervised

# What to summarize Input?

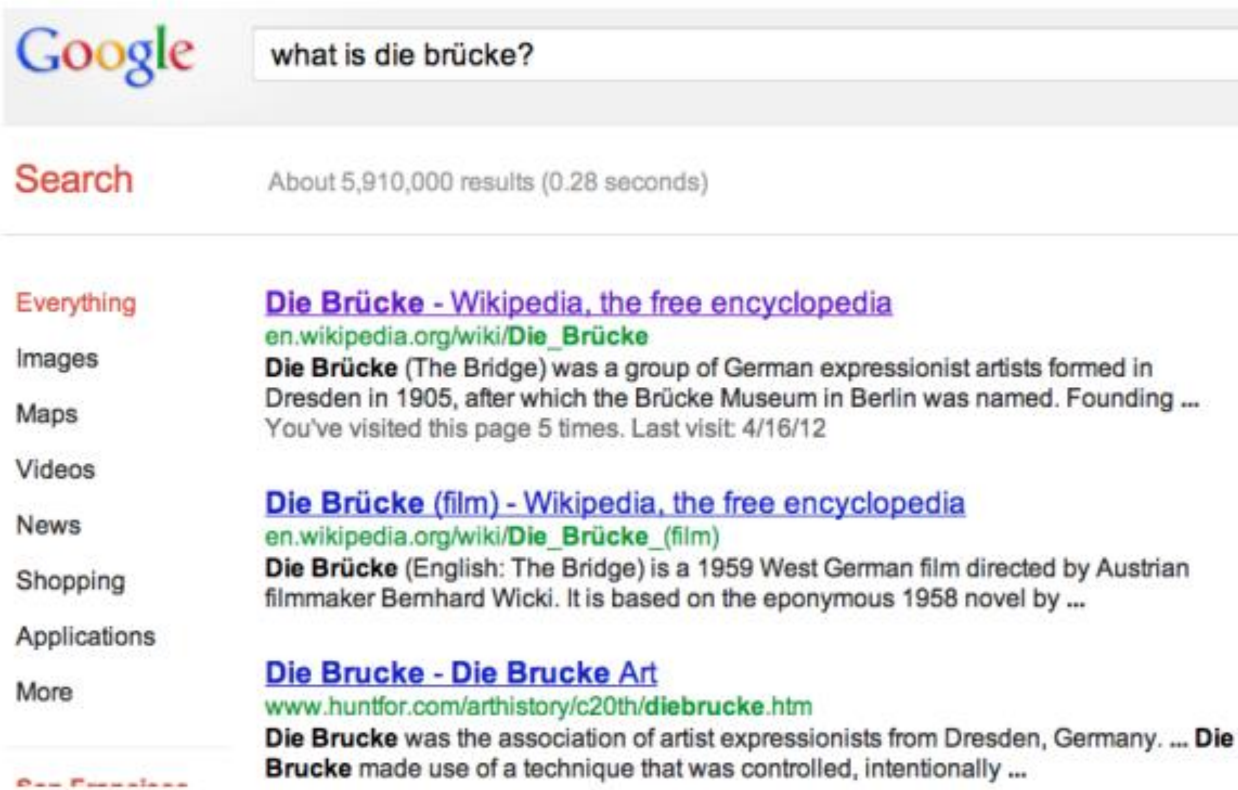
- **Single-document summarization**
  - Given a single document, produce
    - abstract
    - outline
    - headline
- **Multiple-document summarization**
  - Given a group of documents, produce a gist of the content:
    - a series of news stories on the same event
    - a set of web pages about some topic or question

# Type of Summarization

- Generic summarization:
  - Summarize the content of a document
- Query-focused summarization:
  - summarize a document with respect to an information need expressed in a user query.
  - a kind of complex question answering:
    - Answer a question by summarizing a document that has the information to construct the answer

# Summarization for Question Answering Snippets

- **Create snippets** summarizing a web page for a query
  - Google: 156 characters (about 26 words) plus title and link



# Summarization for Question Answering Multiple Documents

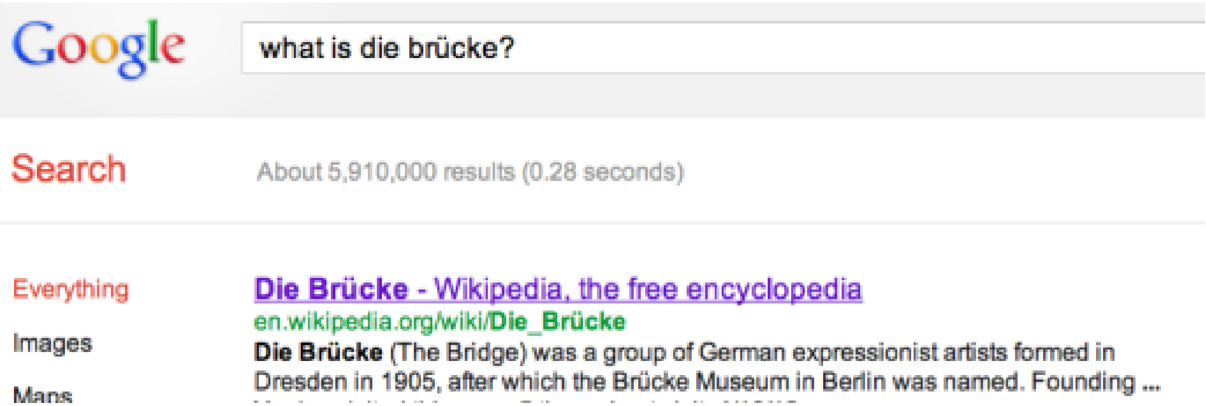
Create answers to complex questions summarizing multiple documents.

- Instead of giving a snippet for each document
- Create a cohesive answer that combines information from each document

# Extractive summarization & Abstractive summarization

- Extractive summarization
  - create the summary from phrases or sentences in the source document(s)
- Abstractive summarization
  - express the ideas in the source documents using (at least in part) different words

# Simple baseline take the first sentence




# Die Brücke

From Wikipedia, the free encyclopedia

*For other uses, see Die Brücke (disambiguation).*

**Die Brücke (The Bridge)** was a group of German expressionist artists formed in Dresden in 1905, after which the **Brücke Museum in Berlin** was named. Founding members were Fritz Bleyl, Erich Heckel, Ernst Ludwig Kirchner and Karl Schmidt-Rottluff. Later members were Emil Nolde, Max Pechstein and Otto Mueller. The seminal group had a major impact on the evolution of modern art in the 20th century and the creation of expressionism.[1]

Die Brücke is sometimes compared to the Fauves. Both movements shared interests in primitivist art. Both

# Query focused summary

Was cast-metal movable type invented in korea?

About 591,000 results (0.14 seconds)

**Movable type** - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Movable_type**
Jump to **Metal movable type**: Transition from wood type to **metal** type occurred in 1234 ... The following description of the **Korean** font **casting** ... In the early fifteenth century, however, the **Koreans invented** a form of **movable type** that has ...

History of printing in East Asia - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/History_of_printing_in_East_Asia
The following description of the **Korean** font **casting** process was recorded by the ... While **metal movable type** printing was **invented in Korea** and the oldest ...

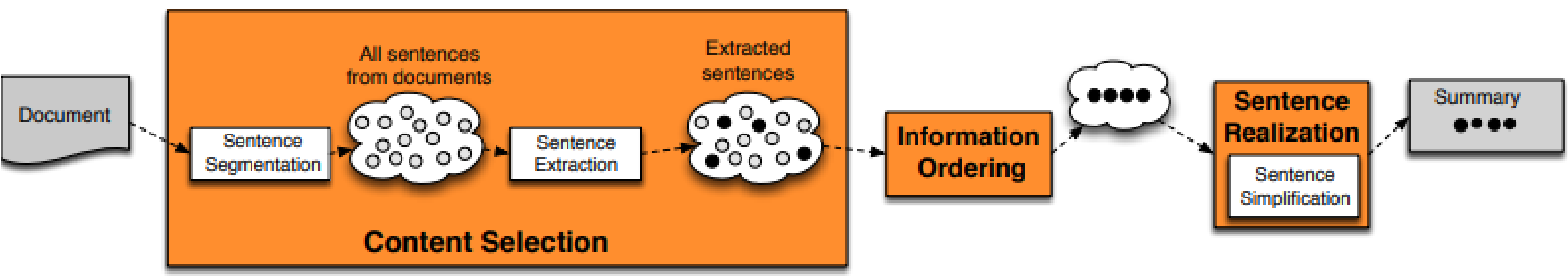
**Korea**, 1000–1400 A.D. | Heilbrunn Timeline of Art History | The ...
www.metmuseum.org/toah/ht/?period=07&region=eak
The **invention** and use of **cast-metal movable type** in **Korea** in the early thirteenth century predates by two centuries Gutenberg's **invention** of metal **movable type** ...

# Summarization Three Stages

1. content selection: choose sentences to extract from the document

2. information ordering: choose an order to place them in the summary

3. sentence realization: clean up the sentence



# Stage 1: Content Selection

# Frequency as indicator of importance

The topic of a document will be repeated many times

In multi-document summarization, important content is repeated in different sources

# Greedy frequency method

Compute word probability from input

Compute sentence weight as function of word probability

Pick best sentence

# *Unsupervised content selection; Luhn (1958)*

### *Intuition*

Choose sentences that have salient or informative words

### *Two approaches to define salient words*

- *tf-idf:* weigh each word $w_i$ in document $j$ by tf-idf

$$weight(w_i) = tf_{ij} \times idf_i$$

- *Topic signatures:* choose a smaller set of salient words, specific to that domain

$$weight(w_i) = 1 \text{ if } w_i \text{ is a specific term (use mutual information)}$$

### *Weighing a sentence*

$$weight(s) = \frac{1}{|S|} \sum_{w \in S} weight(w)$$

# Simple tf*idf

$$w_{ik} = tf_{ik} * \log(N / n_k)$$

$T_k$ = term $k$ in document $D_i$

$tf_{ik}$ = frequency of term $T_k$ in document $D_i$

$idf_k$ = inverse document frequency of term $T_k$ in $C$

$N$ = total number of documents in the collection $C$

$n_k$ = the number of documents in $C$ that contain $T_k$

$$idf_k = \log\left(\frac{N}{n_k}\right)$$

# Using graph representations

Nodes

- Sentences
- Discourse entities

Arcs

- Between similar sentences
- Between related entities

# Using graph representations

## *LexRank: A Graph-based approach*

*Text Document*

Computation is a process following a well defined model ...
A computation can be seen as a purely physical phenomena ...
...

processing →

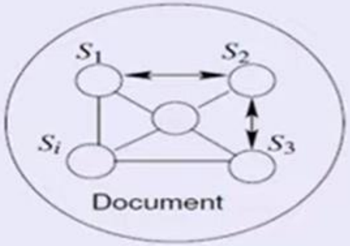
$S_1 \rightarrow \{(computation, 0.1), (process, 0.15), \ldots\}$
$S_2 \rightarrow \{(computation, 0.1), (seen, 0.05), \ldots\}$
$S_3 \rightarrow \ldots$

Machine-readable format

Document Representation

*Underlying Hypothesis*

Sentences that convey the theme of the document are more similar to each other



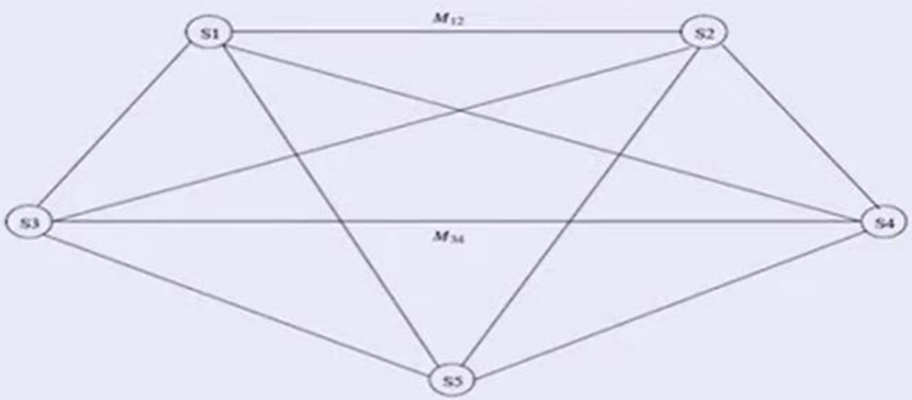

Finding the most salient sentences

# Using graph representations

## Sentence Centrality Measure

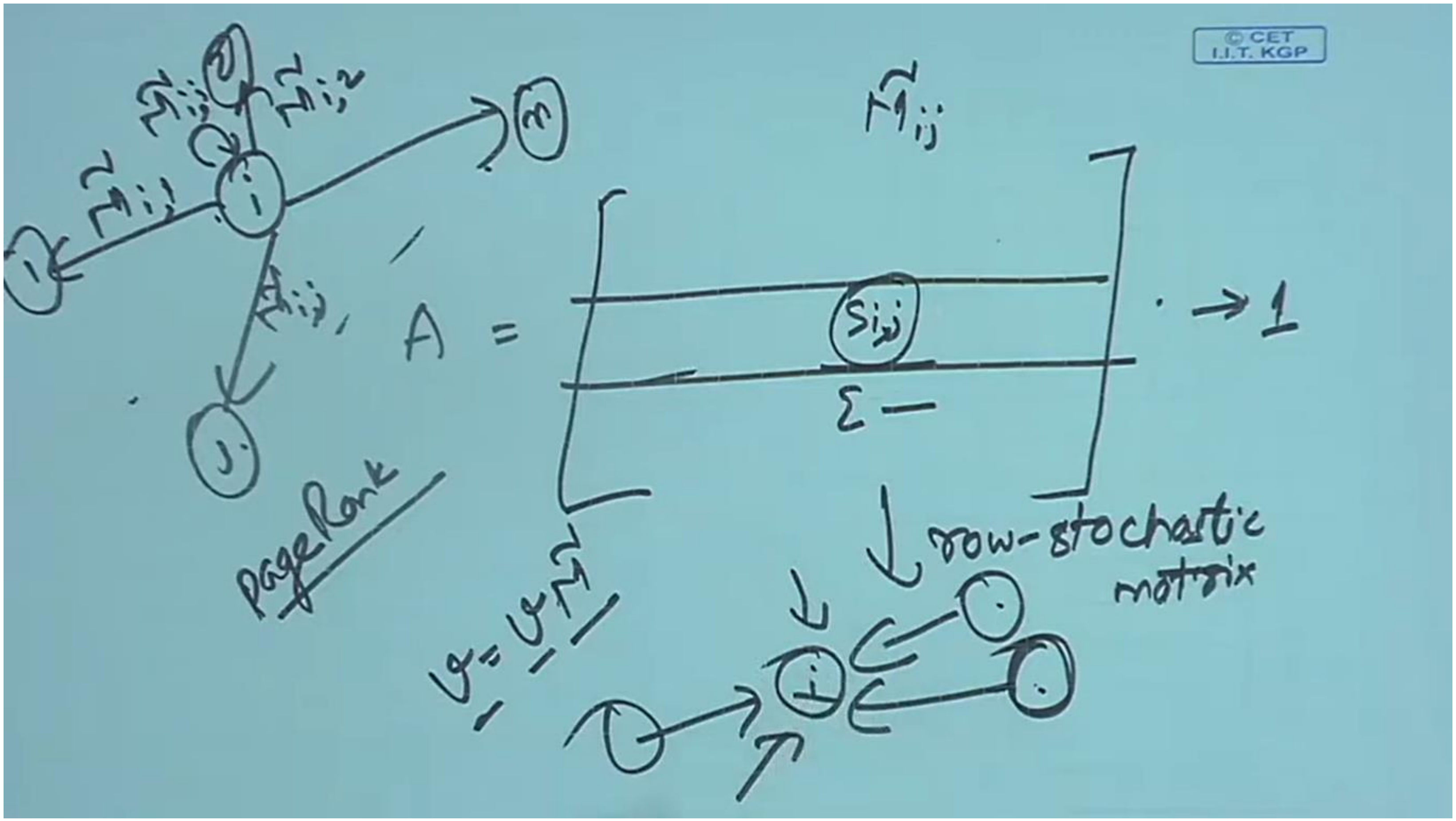
### Finding the most salient sentences

PageRank based algorithm is used to compute the sentence centrality vector $I$.




$$\tilde{M} = \begin{bmatrix} 0.0 & 0.5 & 0.0 & 0.4 & 0.1 \\ 0.5 & 0.0 & 0.5 & 0.0 & 0.0 \\ 0.0 & 0.5 & 0.0 & 0.5 & 0.0 \\ 0.4 & 0.0 & 0.4 & 0.0 & 0.2 \\ 0.3 & 0.0 & 0.0 & 0.7 & 0.0 \end{bmatrix}$$

# Using graph representations




https://www.youtube.com/watch?v=1XBOK-l8Gc8&t=133s

# Supervised Content Selection

- Given:
  - a labeled training set of good summaries for each document
- Align:
  - the sentences in the document with sentences in the summary
- Extract features
  - position (first sentence?)
  - length of sentence
  - word informativeness, cue phrases
  - cohesion
- Train
  - a binary classifier (put sentence in summary? yes or no)

- Problems:
  - hard to get labeled training data
  - alignment difficult
  - performance not better than unsupervised algorithms
- So in practice:
  - **Unsupervised content selection is more common**

# How to deal with redundancy?

Author JK Rowling has won her legal battle in a New York court to get an unofficial Harry Potter encyclopaedia banned from publication.

A U.S. federal judge in Manhattan has sided with author J.K. Rowling and ruled against the publication of a Harry Potter encyclopedia created by a fan of the book series.

– Shallow techniques not likely to work well

# Global optimization for content selection

What is the best summary? vs What is the best sentence?

Form all summaries and choose the best

- What is the problem with this approach?

# MMR:Choosing informative yet non redundant sentences



One of many ways to combine the intuitions of MMR:

1. Score each sentence based on MMR(including query words)
2. Include the sentence with highest score in the summary.
3. Iteratively add into the summary high scoring sentences that are not redundant with summary so far

# Maximal Marginal Relevance MMR

- An iterative method for content selection from multiple documents
- Iteratively (greedily) choose the best sentence to insert in the summary/answer so far:
  - Relevant: Maximally relevant to the user's query
    - high cosine similarity to the query
  - Novel: Minimally redundant with the summary/answer so far
    - low cosine similarity to the summary

$$\hat{s}_{MMR} = \max_{s \in D} \lambda sim(s,Q) - (1\text{-}\lambda) \max_{s \in S} sim(s,S)$$

- Stop when desired length

# Optimization based approach for summarization



- Let us define document $D$ with $t_n$ textual units

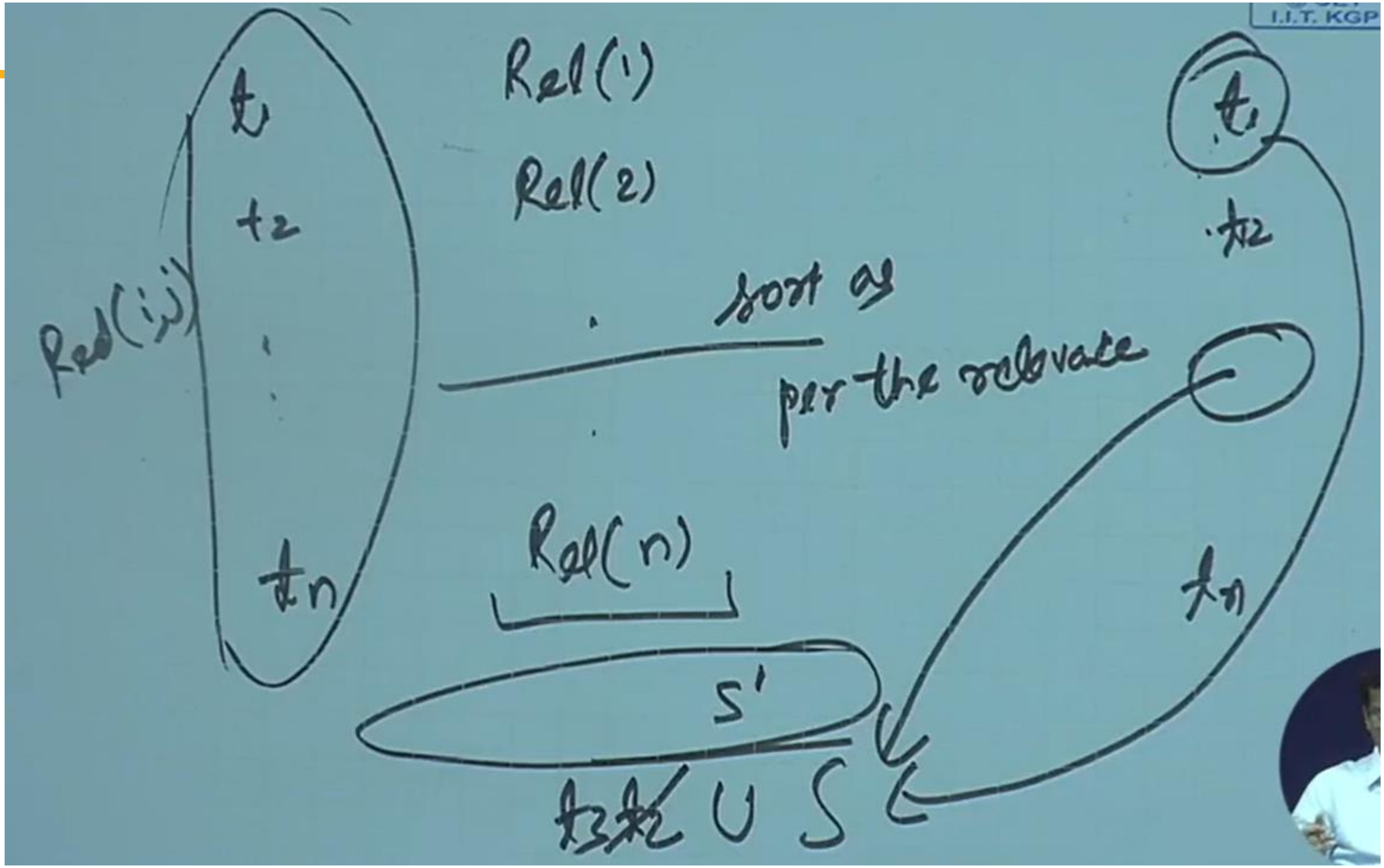
$$D = t_1, t_2, \ldots, t_{n-1}, t_n$$

- Let $Rel(i)$ be the relevance of $t_i$ to be in the summary
- Let $Red(i,j)$ be the redundancy between $t_i$ and $t_j$
- Let $l(i)$ be the length of $t_i$

# Optimization based approach for summarization

- The inference problem is to select a subset $S$ of textual units from $D$ such that summary score of $S$, i.e., $s(S)$, is maximized.
- $S = \arg\max_{S \subseteq D} \left[ \sum_{t_i \in S} Rel(i) - \sum_{t_i, t_j \in S, i<j} Red(i,j) \right]$

  such that $\sum_{t_i \in S} l(i) \leq K$, where $k$ denotes the maximum length of the summary


I.I.T. KGP
Rel(1)
Rel(2)
Rel(n)
sort as per the relevance
Rel(i,j)
t1
t2
tn
S'
U S

# Algorithm

1. Sort $D$ so that $Rel(i) > Rel(i+1) \forall i$
2. $S = \{t_1\}$
3. while $\sum_{t_i \in S} l(i) < K$
4. $\quad t_j = \arg\max_{t_j \in D-S} s(S \cup \{t_j\})$
5. $\quad S = S \cup \{t_j\}$
6. return $S$

# Stage 2: Information Ordering

# Information ordering

In what order to present the selected sentences?

- An article with permuted sentences will not be easy to understand

Very important for multi-document summarization

- Sentences coming from different documents

# Information Ordering

- Chronological ordering:
  - Order sentences by the date of the document (for summarizing news) (Barzilay, Elhadad, and McKeown 2002)
- Coherence:
  - Choose orderings that make neighboring sentences similar (by cosine).
  - Choose orderings in which neighboring sentences discuss the same entity (Barzilay and Lapata 2007)
- Topical ordering
  - Learn the ordering of topics in the source documents

# Domain specific answering: Information Extraction method

- a good biography of a person contains:
  - a person's birth/death, fame factor, education, nationality and so on
- a good definition contains:
  - genus or hypernym
  - Hajj is a type of ritual
- a medical answer about a drug's use contains:
  - • the problem (the medical condition),
  - • the intervention (the drug or procedure), and
  - • the outcome (the result of the study).

# Information that should be in the answer for 3 kinds of questions

| | |
|---|---|
| **Definition** | |
| **genus** | The Hajj is a type of ritual |
| **species** | the annual hajj begins in the twelfth month of the Islamic year |
| **synonym** | The Hajj, or Pilgrimage to Mecca, is the central duty of Islam |
| **subtype** | Qiran, Tamattu', and Ifrad are three different types of Hajj |
| **Biography** | |
| **dates** | was assassinated on April 4, 1968 |
| **nationality** | was born in Atlanta, Georgia |
| **education** | entered Boston University as a doctoral student |
| **Drug efficacy** | |
| **population** | 37 otherwise healthy children aged 2 to 12 years |
| **problem** | acute, intercurrent, febrile illness |
| **intervention** | acetaminophen (10 mg/kg) |
| **outcome** | ibuprofen provided greater temperature decrement and longer duration of antipyresis than acetaminophen when the two drugs were administered in approximately equal doses |

# Answering harder questions: Query focused multi-document summarization

- The (bottom up) snippet method
  - • Find a set of relevant documents
  - • Extract informative sentences from the documents
  - • Order and modify the sentences into an answer
- The (top down) information extraction method
  - build specific answers for different question types:
    - definition questions
    - biography questions
    - certain medical questions

# Definition questions

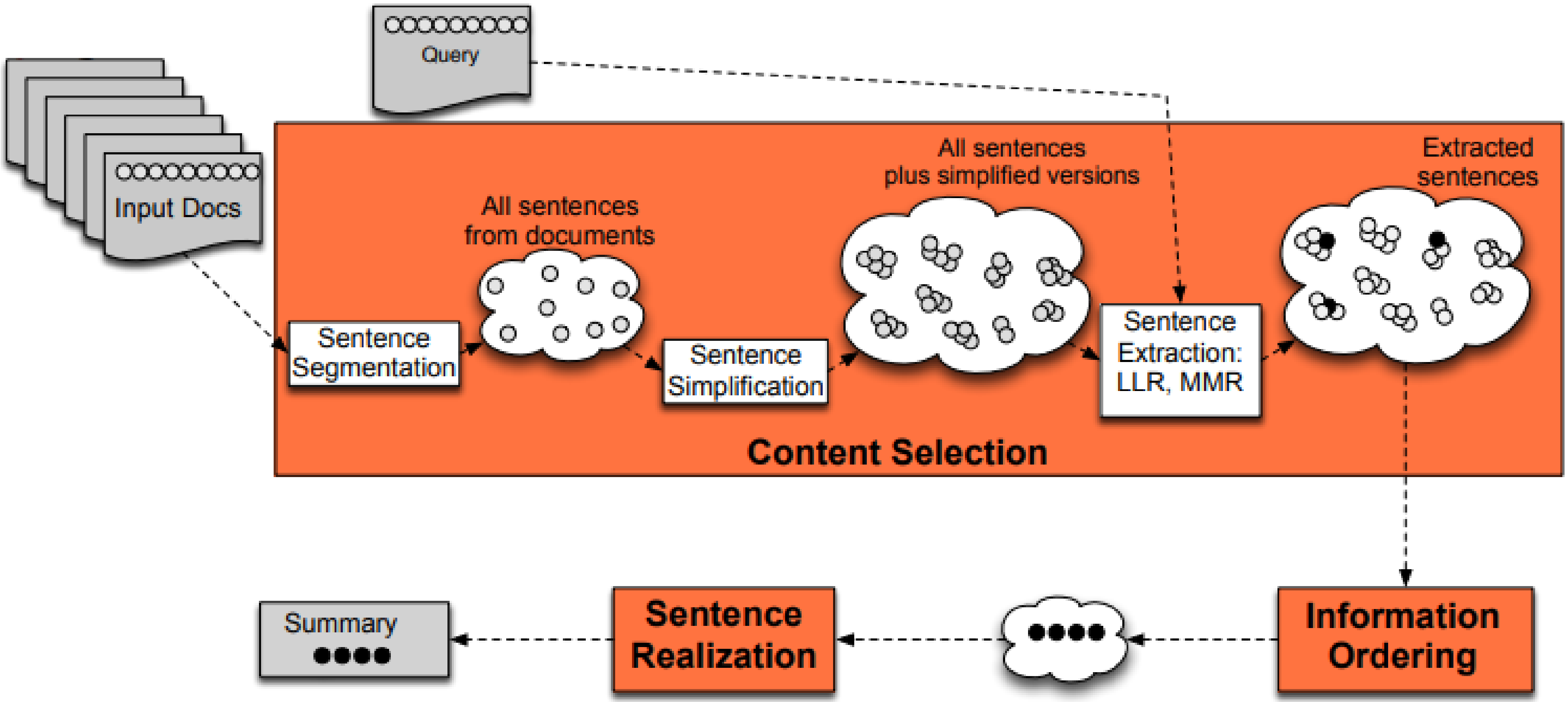
Q: What is water spinach?

A:Water spinach (ipomoea aquatica) is a semiaquatic leafy green plant with long hollow stems and spear or heart shaped leaves, widely grown throughout Asia as a leaf vegetable. The leaves and stems are often eaten fried flavored with salt or in soups. Other common names include morning glory vegetable, kangkong (Malay), It is not related to spinach, but is closely related to sweet potato and convolvulus.

# Complex Questions

1. How is compost made and used for gardening (including different types of compost, their uses, origins and benefits)?
2. What causes train wrecks and what can be done to prevent them?
3. Where have poachers endangered wildlife, what wildlife has been endangered and what steps have been taken to prevent poaching?
4. What has been the human toll in death or injury of tropical storms in recent years?

# Query Focused Multi Document Summarization

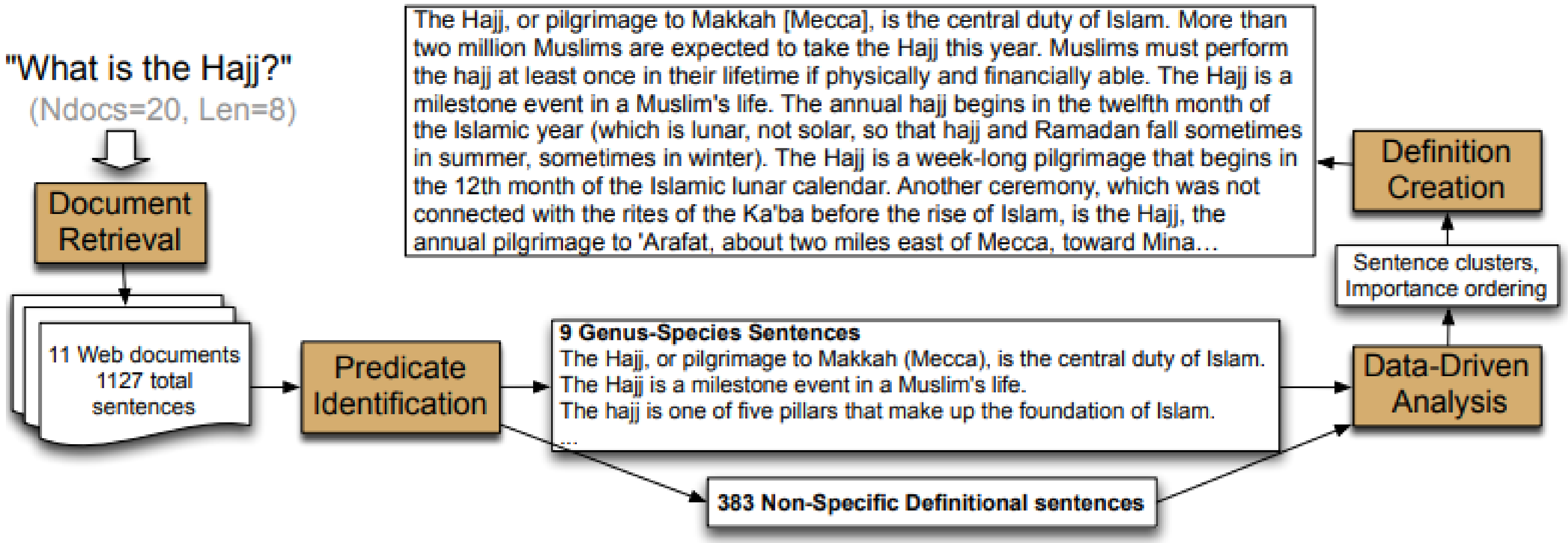


# Simplifying sentences

Simplest method: parse sentences, use rules to decide which modifiers to prune (more recently a wide variety of machine-learning methods)

| | |
|---|---|
| **appositives** | Rajam, ~~28, an artist who was living at the time in Philadelphia~~, found the inspiration in the back of city magazines. |
| **attribution clauses** | Rebels agreed to talks with government officials, ~~international observers said Tuesday.~~ |
| **PPs without named entities** | The commercial fishing restrictions in Washington will not be lifted unless the salmon population increases [~~PP to a sustainable number~~]] |
| **initial adverbials** | "~~For example~~", "~~On the other hand~~", "~~As a matter of fact~~", "~~At this point~~" |

# Architecture for complex question answering: definition questions



# Automatic summary edits

- Some expressions might not be appropriate in the new context
  - References:
    - he
    - Putin
    - Russian Prime Minister Vladimir Putin
  - Discourse connectives
    - However, moreover, subsequently
- Requires more sophisticated NLP techniques

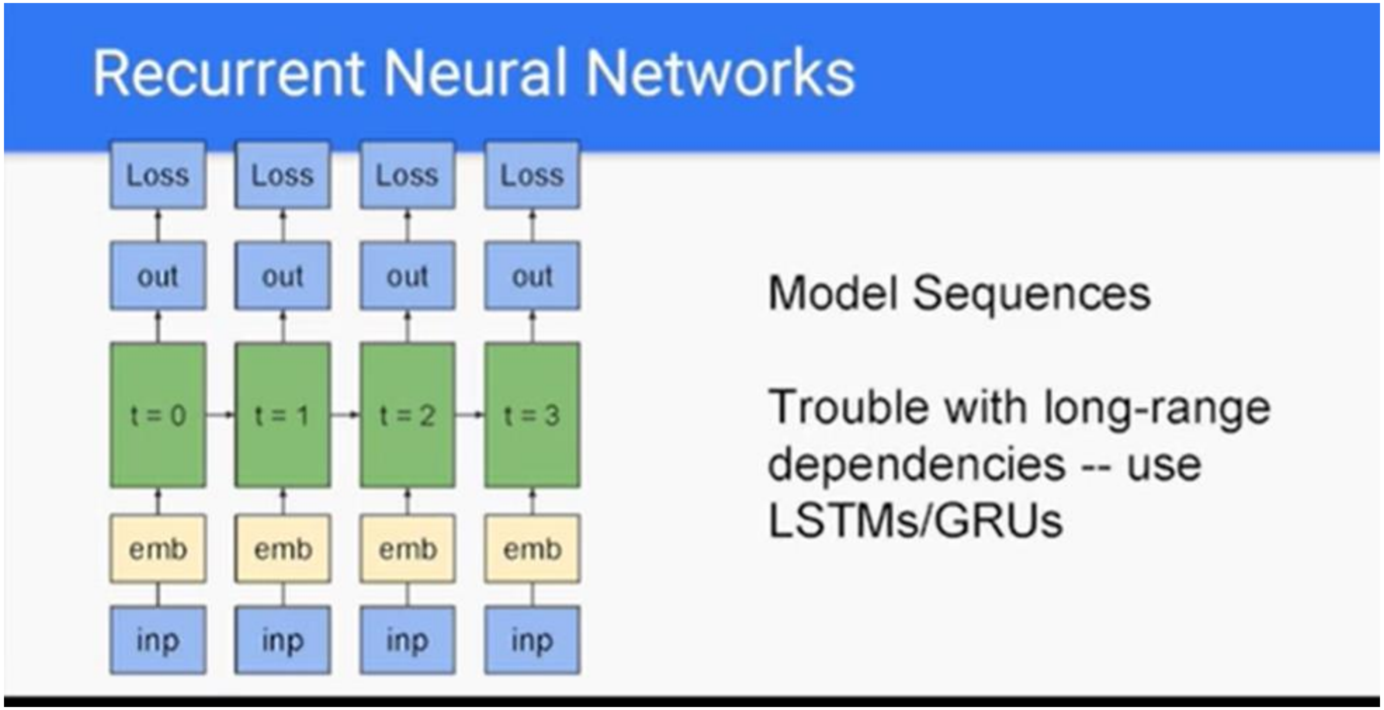
# Before

Pinochet was placed under arrest in London Friday by British police acting on a warrant issued by a Spanish judge. Pinochet has immunity from prosecution in Chile as a senator-for-life under a new constitution that his government crafted. Pinochet was detained in the London clinic while recovering from back surgery.

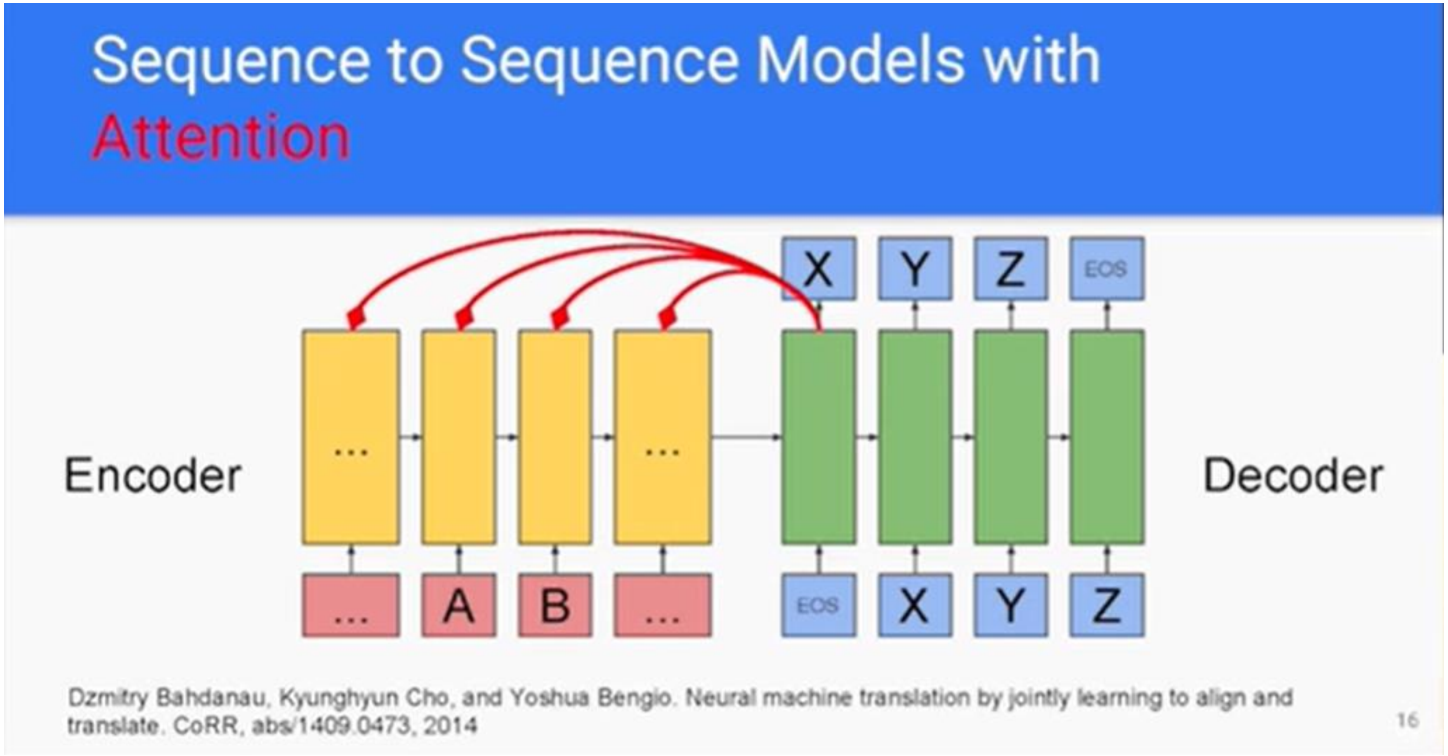
# After

Gen. Augusto Pinochet, the former Chilean dictator, was placed under arrest in London Friday by British police acting on a warrant issued by a Spanish judge. Pinochet has immunity from prosecution in Chile as a senator-for-life under a new constitution that his government crafted. Pinochet was detained in the London clinic while recovering from back surgery.

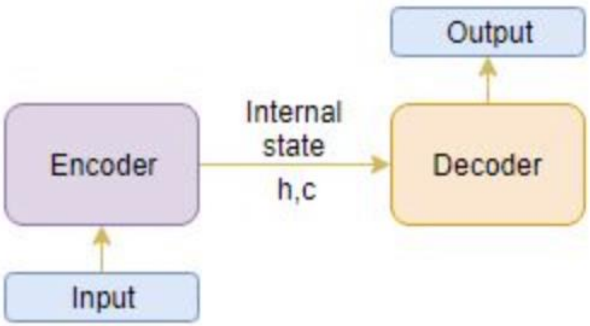
# Neural Text Summarization

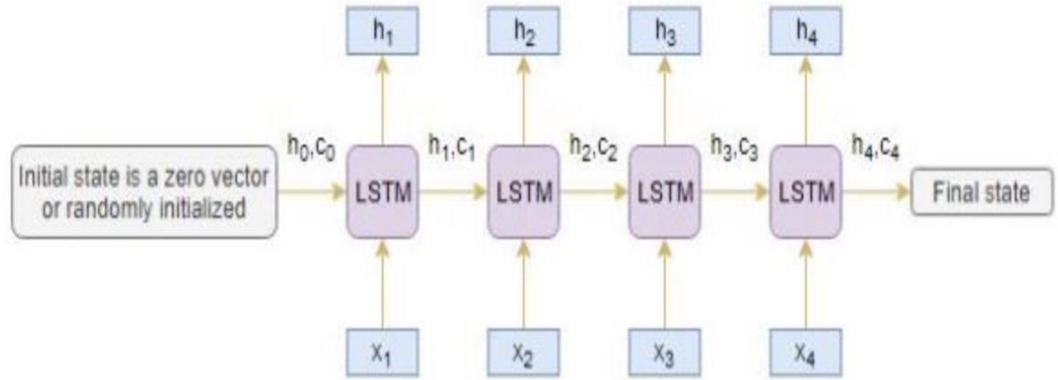


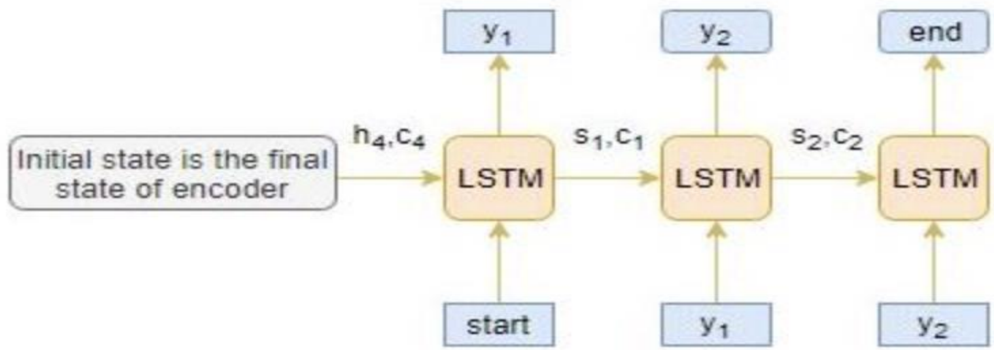
# Neural Text Summarization



# Neural Text Summarization

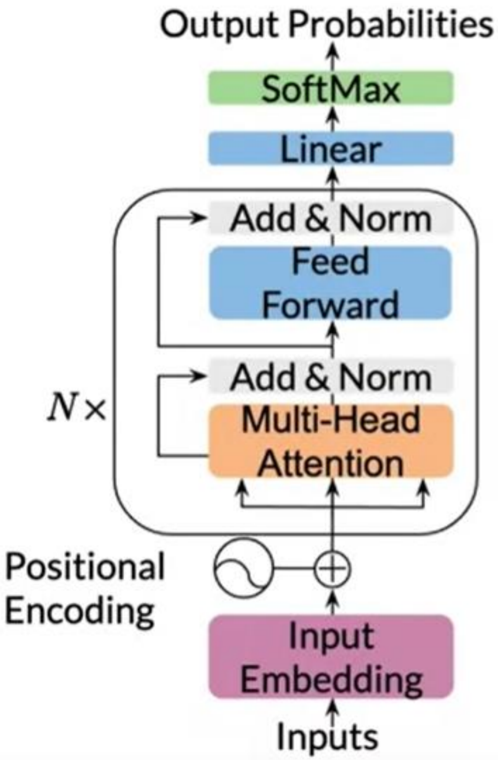


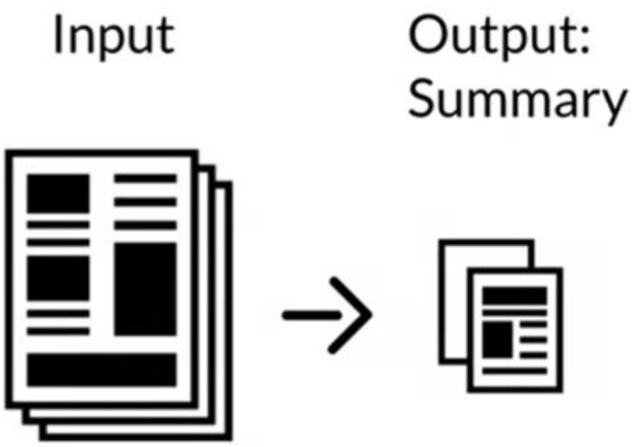



# Neural Text Summarization

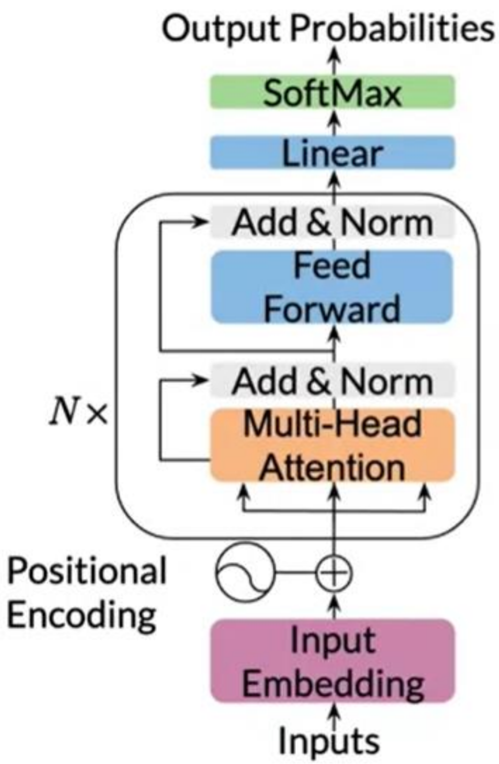
## Transformer for summarization



# Neural Text Summarization

## Technical details for data processing




**Model Input:**

```
ARTICLE TEXT <EOS> SUMMARY <EOS> <pad> ...
```

**Tokenized version:**

```
[2,3,5,2,1,3,4,7,8,2,5,1,2,3,6,2,1,0,0]
```

Loss weights: 0s until the first <EOS> and then 1 on the start of the summary.

# Approaches Summary

## Generation Way

- `gen-ext` : Extractive Summarization
- `gen-abs` : Abstractive Summarization
- `gen-2stage` Two-stage Summarization (compressive, hybrid)

## Regressive Way

- `regr-auto` : Autoregressive Decoder (Pointer network)
- `regr-nonauto` : Non-autoregressive Decoder (Sequence labeling)

## Supervision

- `sup-sup` : Supervised Learning
- `sup-weak` (implies `sup-sup` ): Weakly Supervised Learning
- `sup-unsup` : Unsupervised Learning

## Task Settings

rich of task settings!

- `task-single` : Single-document Summarization
- `task-multi` : Multi-document Summarization
- `task-senCompre:` Sentence Compression
- `task-sci` : Scientific Paper
- `task-multimodal` : Multi-modal Summarization
- `task-aspect` : Aspect-based Summarization
- `task-opinion` : Opinion Summarization
- `task-questoin` : Question-based Summarization

## Architecture (Mechanism)

- `arch-rnn` : Recurrent Neural Networks (LSTM, GRU)
- `arch-cnn` : Convolutional Neural Networks (CNN)
- `arch-transformer` : Transformer
- `arch-graph` : Graph Neural Networks or Statistic Graph Models
- `arch-gnn` : Graph Neural Networks
- `arch-att` : Attention Mechanism
- `arch-pointer` : Pointer Layer
- `arch-coverage` : Coverage Mechanism

## Training

- `train-multitask` : Multi-task Learning
- `train-multilingual` : Multi-lingual Learning
- `train-multimodal` : Multi-modal Learning
- `train-auxiliary` : Joint Training
- `train-transfer` : Cross-domain Learning, Transfer Learning, Domain Adaptation
- `train-active` : Active Learning, Boostrapping
- `train-adver` : Adversarial Learning
- `train-template` : Template-based Summarization
- `train-augment` : Data Augmentation
- `train-curriculum` : Curriculum Learning
- `train-lowresource` : Low-resource Summarization
- `train-retrieval` : Retrieval-based Summarization
- `train-meta` : Meta-learning

## Pre-trained Models

- `pre-word2vec` : word2vec
- `pre-glove` : GLoVe
- `pre-bert` : BERT

# Evaluating Summaries: ROUGE

ROUGE (Recall Oriented Understudy for Gisting Evaluation)

- Intrinsic metric for atomically evaluating summaries
- Based on BLEU (a metric used for machine translation)
- Not as good as human evaluation ("Did this answer the user's question?")
- But much more convenient

# ROUGE-2

Given a document D, and an automatic summary X:

1. Have N humans produce a set of reference summaries of D
2. Run system, giving automatic summary X
3. What percentage of the bigrams from the reference summaries appear in X?

$$ROUGE-2 = \frac{\sum_{s \in \{\text{RefSummaries}\}} \sum_{\text{bigrams } i \in S} \min(count(i,X), count(i,S))}{\sum_{s \in \{\text{RefSummaries}\}} \sum_{\text{bigrams } i \in S} count(i,S)}$$

# ROUGE-2 Example

Q: “What is water spinach?”

Human 1: Water spinach is a green leafy vegetable grown in the tropics.

Human 2: Water spinach is a semi-aquatic tropical plant grown as a vegetable.

Human 3: Water spinach is a commonly eaten leaf vegetable of Asia.

- System answer: Water spinach is a leaf vegetable commonly eaten in tropical areas of Asia.

Rouge-2 score= $\dfrac{3+3+6}{10+9+9} = 12/28 = .43$

# References


- Speech and Language processing An introduction to Natural Language Processing, Computational Linguistics and speech Recognition by Daniel Jurafsky and James H. Martin[3rd edition] Chapter 21
- https://www.youtube.com/watch?v=9PoKellNrBc
- https://www.youtube.com/watch?v=x9h5vJpkV_8
- http://www.infocobuild.com/education/audio-video-courses/computer-science/NaturalLanguageProcessing-IIT-Kharagpur/lecture-52.html
- https://harvard-iacs.github.io/CS287/lectures/14_Summarization.pdf
- http://demo.clab.cs.cmu.edu/algo4nlp19/slides/summarization.pdf
- https://people.engr.tamu.edu/huangrh/Fall16/l22_text_summarization.pdf

- https://vimeo.com/193652155
- https://www.turing.com/kb/5-powerful-text-summarization-techniques-in-python