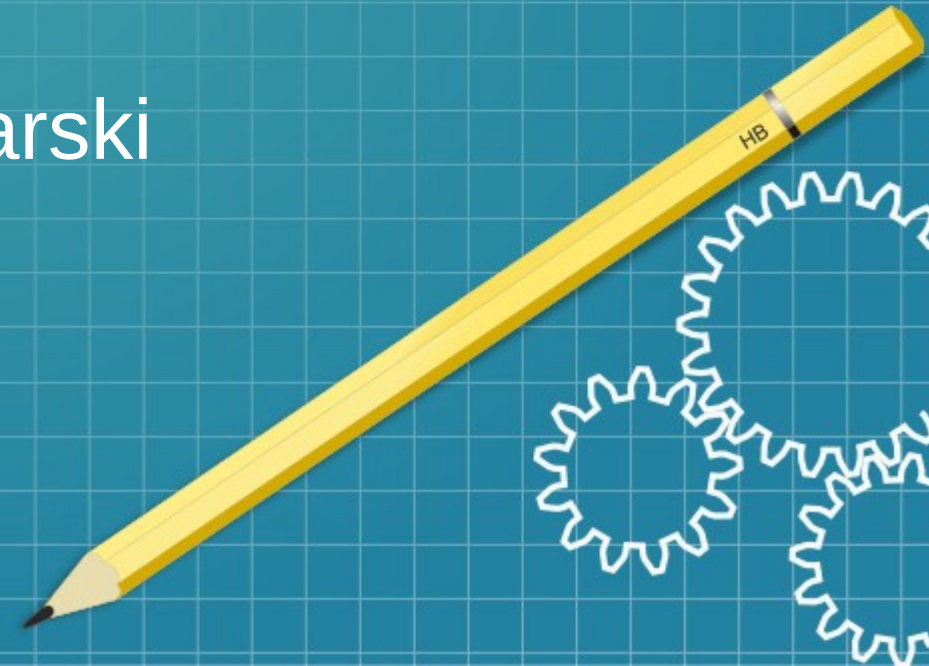
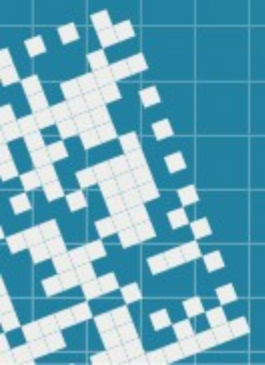


# Semantics analysis

Tomasz Garski



# Extractive Summarization

A yellow pencil with a black eraser and a pink eraser are positioned in the top right corner of the slide.

- Extractive summarization means identifying important sections of the text and generating them verbatim producing a subset of the sentences from the original text.

# Extractive Summarization



- 1) Construction of an intermediate representation of the input text
- 
- There are two types of representation-based approaches: topic representation and indicator representation. Topic representation transforms the text into an intermediate representation and interpret the topic(s) discussed in the text. The techniques used for this differ in terms of their complexity, and are divided into frequency-driven approaches, topic word approaches, latent semantic analysis and Bayesian topic models. Indicator representation describes every sentence as a list of formal features (indicators) of importance such as sentence length, position in the document, having certain phrases, etc.

# Extractive Summarization



- 2) Scoring the sentences based on the representation
- 
- When the intermediate representation is generated, an importance score is assigned to each sentence. In topic representation approaches, the score of a sentence represents how well the sentence explains some of the most important topics of the text. In indicator representation, the score is computed by aggregating the evidence from different weighted indicators.



# Extractive Summarization



- 3) Selection of a summary comprising of a number of sentences
- 
- The summarizer system selects the top  $k$  most important sentences to produce a summary. Some approaches use greedy algorithms to select the important sentences and some approaches may convert the selection of sentences into an optimization problem where a collection of sentences is chosen, considering the constraint that it should maximize overall importance and coherency and minimize the redundancy.

# Extractive Summarization

A yellow pencil and a pink eraser are positioned in the top right corner of the slide, appearing to be on the paper.

- Input document → sentences similarity → weight sentences → select sentences with higher rank.

# Abstractive Summarization

A yellow pencil with a pink eraser is positioned in the top right corner of the slide, pointing towards the title.

- Tree-based methods
- The central idea of this bunch of methods is using a dependency tree that represents the text or the contents of a document. At the same time, the algorithms of content selection vary significantly from theme intersection to different algorithms are used for content choice for outline e.g. algorithmic program or native alignment try across of parsed sentences. The outline is generated either with the help of a language generator or an associate degree algorithm.

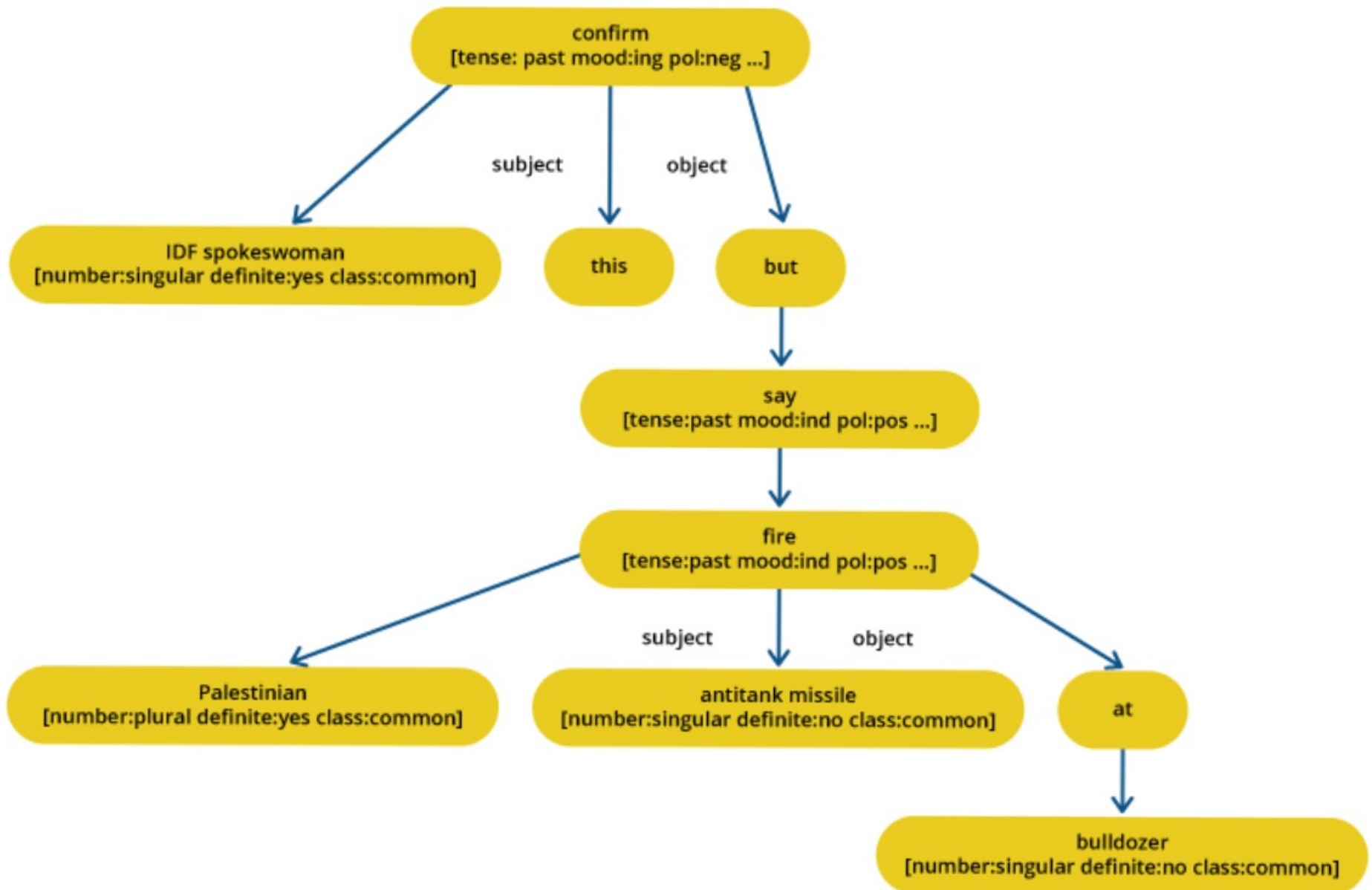
# Abstractive Summarization

A yellow pencil is positioned diagonally in the top right corner, pointing towards the bottom left. Below the pencil's tip is a small, rectangular pink eraser.

- An example of such approach is sentence fusion—the algorithm which processes multiple documents, identifies common information by aligning syntactic trees of input sentences, incorporating paraphrasing information, then matches subsets of the subtrees through bottom-up local multisequence alignment, combines fragments through construction of a fusion lattice encompassing the resulting alignment and transforms the lattice into a sentence using a language model. The approach therefore combines statistical techniques, such as local, multisequence alignment and language modeling, with linguistic representations automatically derived from input documents.



# Abstractive Summarization

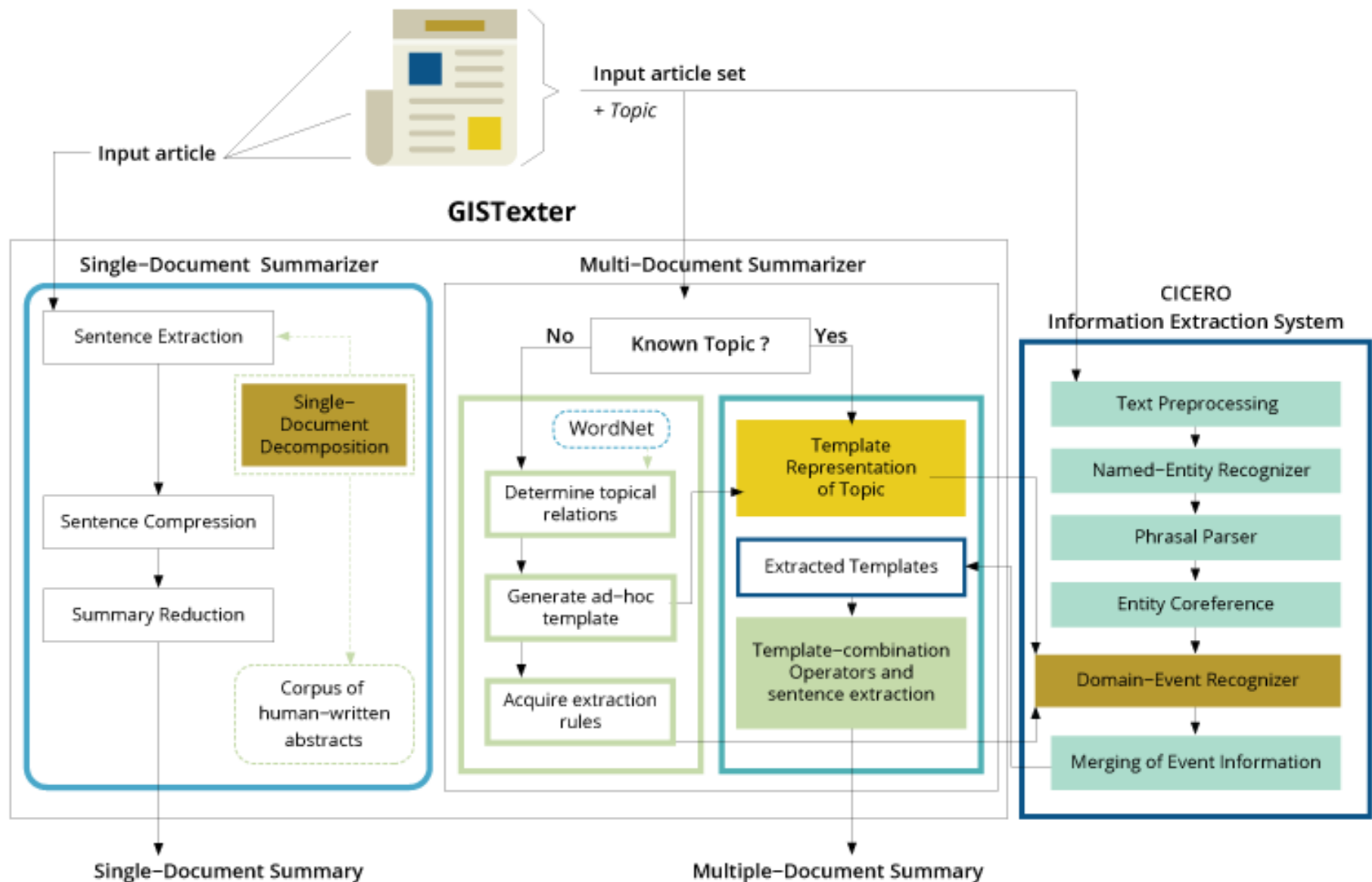


# Abstractive Summarization



- Template-based methods
- In this method, a full document is represented using a certain guide. Linguistic patterns or extraction rules are matched to spot text snippets that may be mapped into the guide slots (to form a database). These text snippets serve as the indicators of the outline content. An example of such approach is GISTEXTER, a summarization system that targets the identification of topic-related information in the input document, translates it into database entries and adds sentences from this database to ad hoc summaries.

# Abstractive Summarization



# Rogue - evaluation

A yellow pencil is positioned diagonally in the top right corner, pointing towards the bottom left. Below the pencil's tip is a small, rectangular pink eraser.

- ROUGE stands for Recall-Oriented Understudy for Gisting Evaluation. It is essentially of a set of metrics for evaluating automatic summarization of texts as well as machine translation. It works by comparing an automatically produced summary or translation against a set of reference summaries (typically human-produced).

# Rogue - evaluation



- Let us say, we have the following system and reference summaries:
  -
- System Summary (what the machine produced):
  - 
  - the cat was found under the bed
  -
- Reference Summary (gold standard - usually by humans) :
  - 
  - the cat was under the bed
  -
- If we consider just the individual words, the number of overlapping words between the system summary and reference summary is 6. This however, does not tell you much as a metric. To get a good quantitative value, we can actually compute the precision and recall using the overlap.



# Rogue - recall



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

$$\text{Recall} = \frac{6}{6} = 1.0$$

# Rogue - precision



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

$$\text{Precision} = \frac{6}{7} = 0.86$$

# Rogue - precision



ROUGE-N - measures unigram, bigram, trigram and higher order n-gram overlap

ROUGE-L - measures longest matching sequence of words using LCS. An advantage of using LCS is that it does not require consecutive matches but in-sequence matches that reflect sentence level word order. Since it automatically includes longest in-sequence common n-grams, you don't need a predefined n-gram length.

ROUGE-S - Is any pair of word in a sentence in order, allowing for arbitrary gaps. This can also be called skip-gram cooccurrence. For example, skip-bigram measures the overlap of word pairs that can have a maximum of two gaps in between words. As an example, for the phrase "cat in the hat" the skip-bigrams would be "cat in, cat the, cat hat, in the, in hat, the hat".

# Query-Based Abstractive Summarization Using Neural Networks

A yellow pencil is positioned diagonally in the top right corner, pointing towards the title. Below it, a pink eraser is also positioned diagonally.

Johan Hasselqvist, Niklas Helmertz, Mikael Kågebäck

(Submitted on 17 Dec 2017)


# Query-Based Abstractive Summarization Using Neural Networks



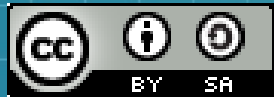
Johan Hasselqvist, Niklas Helmertz, Mikael Kågebäck

(Submitted on 17 Dec 2017)





# Query-based extraction, and multi-document summarization of law regulations



Thank you for your attention.

