Text Summarization

Text Summarization

- **Goal**: produce an abridged version of a text that contains information that is important or relevant to a user.

- **Summarization Applications**
  - **outlines or abstracts** of any document, article, etc
  - **summaries** of email threads
  - **action items** from a meeting
  - **simplifying** text by compressing sentences

What to summarize? Single vs. multiple documents

- **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

Query-focused Summarization & Generic Summarization

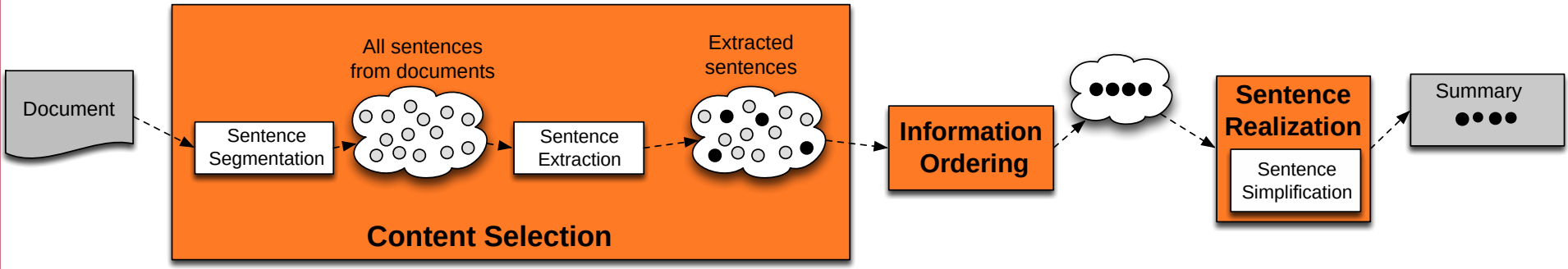
- Generic summarization:
  - Summarize the content of a document
- Query-focused summarization:
  - summarize a document with respect to an information need expressed in a user query.
  - a kind of complex question answering:
    - Answer a question by summarizing a document that has the information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

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

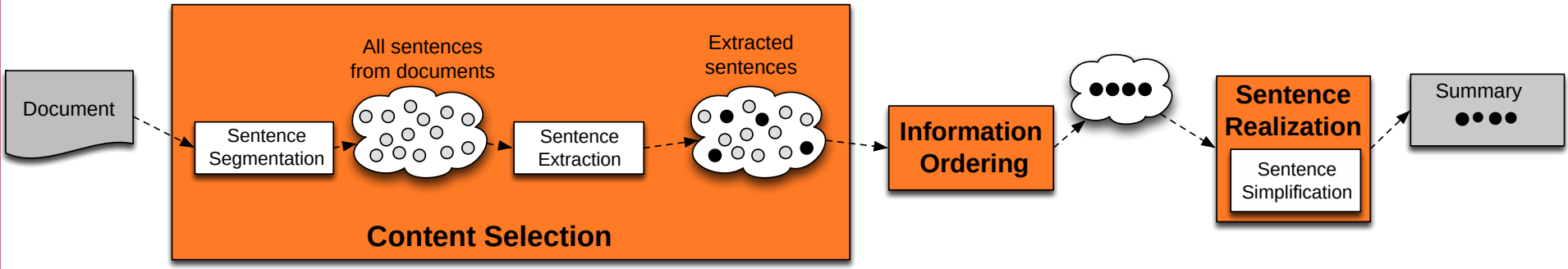
Summarization: Three Stages

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 sentences



Basic Summarization Algorithm

1. content selection: choose sentences to extract from the document
2. information ordering: just use document order
3. sentence realization: keep original sentences



Unsupervised content selection

H. P. Luhn. 1958. The Automatic Creation of Literature Abstracts. IBM Journal of Research and Development. 2:2, 159-165.

- Intuition dating back to Luhn (1958):
  - Choose sentences that have salient or informative words
- Two approaches to defining salient words
  1. tf-idf: weigh each word w_i in document j by tf-idf

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

  2. topic signature: choose a smaller set of salient words
     - mutual information
     - log-likelihood ratio (LLR) Dunning (1993), Lin and Hovy (2000)

$$weight(w_i) = \begin{cases} 1 & \text{if } -2\log \lambda(w_i) > 10 \\ 0 & \text{otherwise} \end{cases}$$

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**

Fourscore and seven years ago our fathers brought forth on this continent a new nation, conceived in liberty, and dedicated to the proposition that all men are created equal. Now we are engaged in a great civil war, testing whether that nation, or any nation so conceived and so dedicated, can long endure. We are met on a great battle- field of that war. We have come to dedicate a portion of that field as a final resting-place for those who here gave their lives that this nation might live. It is altogether fitting and proper that we should do this. But, in a larger sense, we cannot dedicate...we cannot consecrate...we cannot hallow... this ground. The brave men, living and dead, who struggled here, have consecrated it far above our poor power to add or detract. The world will little note nor long remember what we say here, but it can never forget what they did here. It is for us, the living, rather, to be dedicated here to the unfinished work which they who fought here have thus far so nobly advanced. It is rather for us to be here dedicated to the great task remaining before us...that from these honored dead we take increased devotion to that cause for which they gave the last full measure of devotion; that we here highly resolve that these dead shall not have died in vain; that this nation, under God, shall have a new birth of freedom; and that government of the people, by the people, for the people, shall not perish from the earth.

Figure 23.12 The Gettysburg Address. Abraham Lincoln, 1863.

Extract from the Gettysburg Address:

Four score and seven years ago our fathers brought forth upon this continent a new nation, conceived in liberty, and dedicated to the proposition that all men are created equal. Now we are engaged in a great civil war, testing whether that nation can long endure. We are met on a great battlefield of that war. We have come to dedicate a portion of that field. But the brave men, living and dead, who struggled here, have consecrated it far above our poor power to add or detract. From these honored dead we take increased devotion to that cause for which they gave the last full measure of devotion — that government of the people, by the people for the people shall not perish from the earth.

Abstract of the Gettysburg Address:

This speech by Abraham Lincoln commemorates soldiers who laid down their lives in the Battle of Gettysburg. It reminds the troops that it is the future of freedom in America that they are fighting for.

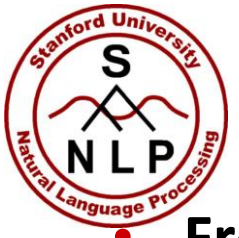
Figure 23.13 An extract versus an abstract from the Gettysburg Address (abstract from Mani (2001)).

Extractive Summarization

- The process of extractive summarizing involves picking the most relevant sentences from an article and systematically organizing them.
- The sentences making up the summary are taken verbatim from the source material.
- Extractive summarization systems revolve around three fundamental operations:
  - **Construction of an intermediate representation of the input text**
    - Topic representation and indicator representation are examples of representation-based methods
  - **Scoring the sentences based on the representation**
    - Each sentence is given a significance score. A sentence's score reflects how effectively it elucidates critical concepts in the text
  - **Selection of a summary comprising several sentences**
    - Choose relevant sentences

Topic Representation Approaches

- **Topic words**
  - Using this method, we can find terms related to the topic in an input document.
  - A sentence's significance can be calculated in two ways:
    - First, as a function of the number of topic signatures it includes
    - Second, as a fraction of the topic signatures it contains.
  - While the first method gives higher scores to longer sentences with more words, the second one measures the density of the topic words.
  - **Topic signature**, a set of **salient** or **signature terms**, each of whose saliency scores is greater than some threshold θ



Topic Representation Approaches

- **Frequency-driven approaches**
  - Words are given relative importance.
  - If the term fits the topic, it gets 1 point; otherwise, it reaches zero.
  - Depending on how they are implemented, the weights might be continuous.
  - Topic representations may be achieved using one of two methods:
    - Word Probability
    - TFIDF (Term Frequency Inverse Document Frequency)

Topic Representation Approaches

- **Frequency-driven approaches**
  - **Word Probability**
    - It only takes a word's frequency to indicate its significance.
    - To calculate the likelihood of a word w, we divide the frequency with which it occurs, f (w), by the total number of words, N.

$$P(w) = \frac{f(w)}{N}$$

    - The average significance of the words in a sentence gives the importance of the sentence when using word probabilities.

Topic Representation Approaches

- **Frequency-driven approaches**
  - **TFIDF (Term Frequency Inverse Document Frequency)**
    - This method is an improvement upon the word probability approach.
    - Here, the weights are determined by using the TF-IDF approach.
    - The Term Frequency Inverse Document Frequency (TFIDF) technique gives less importance to terms that often appear in most documents.
    - The weight of each word w in document d is computed as follows:

$$q(w) = f_d(w) * log\frac{|D|}{f_D(w)}$$

where **fd (w)** is the term frequency of word w in document d, **fD (w**) is the number of documents that contain the word w, and |D| is the number of documents in the collection D.

Indicator Representation Approaches

- **Graph-Based methods**
  - Graph methods represent the documents as a connected graph.
  - Sentences form the graph's vertices, and the edges connecting sentences show the degree to which two sentences are related to one another.
  - One method often used to link two vertices is to assess the degree to which two sentences are similar, and if the degree of similarity is higher than a certain threshold, the vertices are connected. Both outcomes are possible with this graph representation.
  - First, the graph's partitions (sub-graphs) define individual categories of information covered by the documents.
  - The second result is that the document's key sentences have been highlighted.
  - Sentences connected to many other sentences in the partition are possibly the center of the graph and are more likely to be included in the summary..

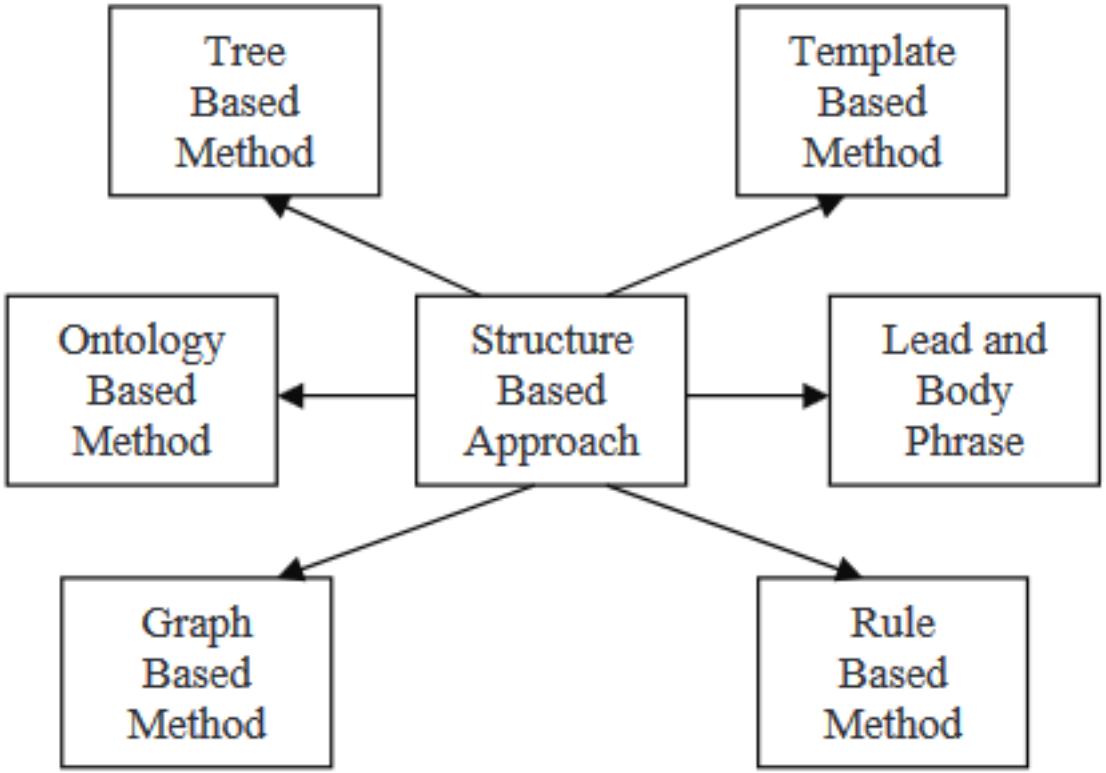
Indicator Representation Approaches

- **Machine Learning**
  - Machine learning techniques see the summarization problem as a classification challenge.
  - Models attempt to categorize sentences into summary and non-summary categories based on their features.
  - We have a training set consisting of documents and human-reviewed extracted summaries from which to train our algorithms.
  - It is often done using Naive Bayes, Decision Tree, or Support Vector Machine.

Abstractive Summarization

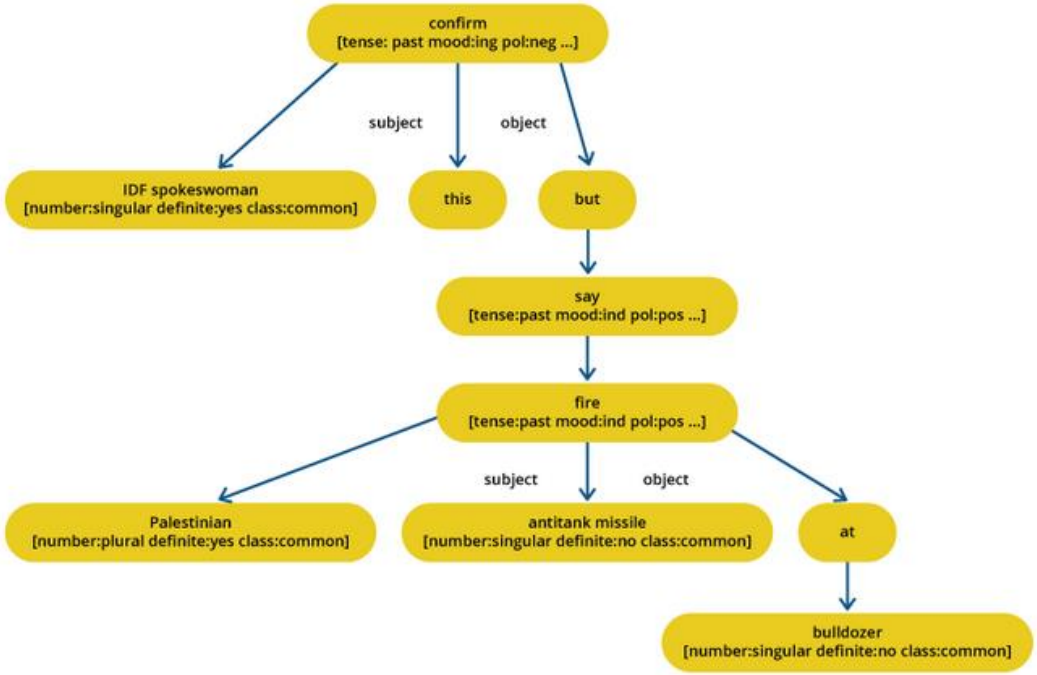
- An abstractive summarizer presents the material in a logical, well-organized, and grammatically sound form.
- A summary's quality can be significantly enhanced by making it more readable or improving its linguistic quality.
- The capacity to create unique sentences that convey vital information from text sources has contributed to the rising appeal of abstractive summarization methods.
- There are two approaches:
  - The Structured based approach
  - Semantic-Based approach

STRUCTURE-BASED APPROACH



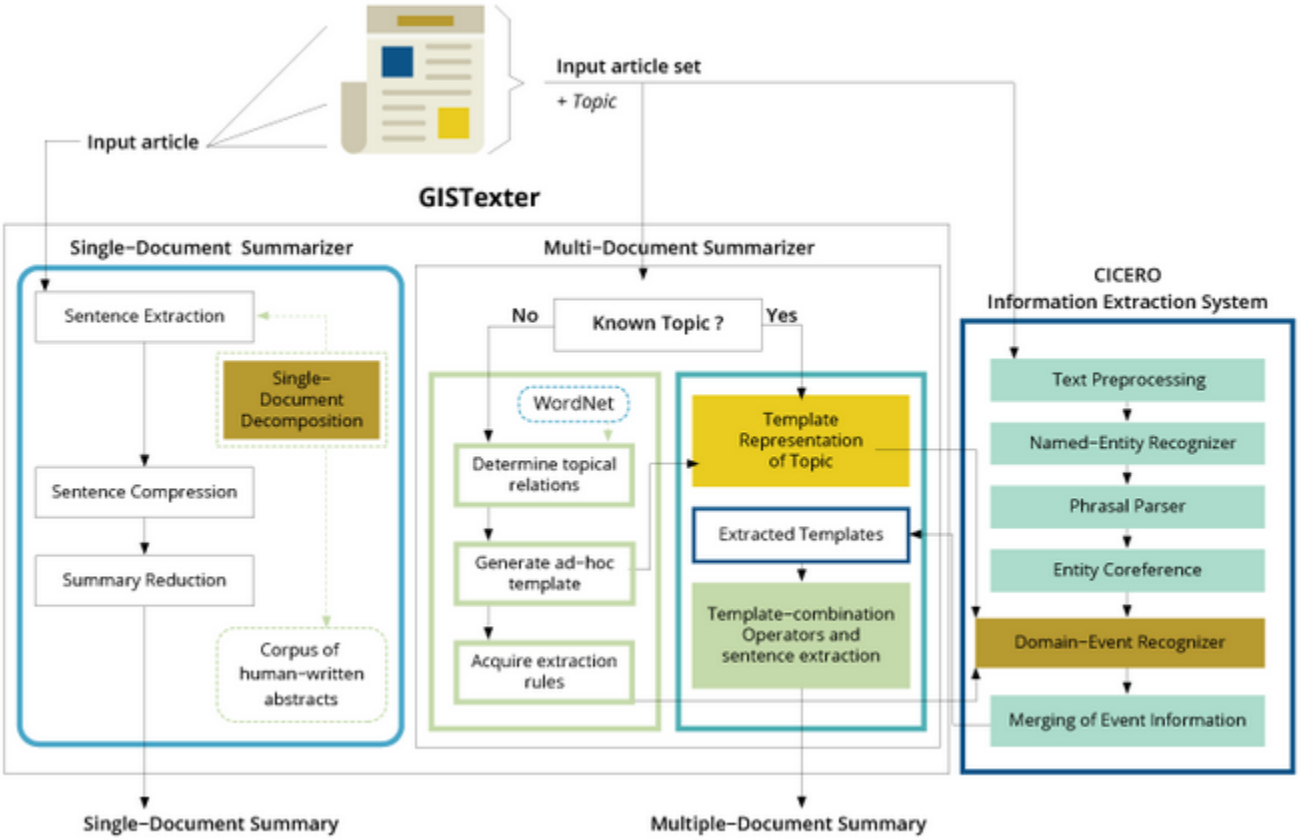
STRUCTURE-BASED APPROACH

- **Tree-based methods**
  - The content of a document is represented as a dependency tree



STRUCTURE-BASED APPROACH

- **Template-based methods**



STRUCTURE-BASED APPROACH

- **Ontology-based methods**
  - Many researchers have attempted to improve summary effectiveness using ontology (knowledge base).
  - Most internet documents have a common domain, meaning they all deal with the same general subject.
  - Ontology is a powerful representation of the unique information structure of each domain.
  - Fuzzy ontology, models uncertainty and accurately describes domain knowledge, to summarize Chinese news

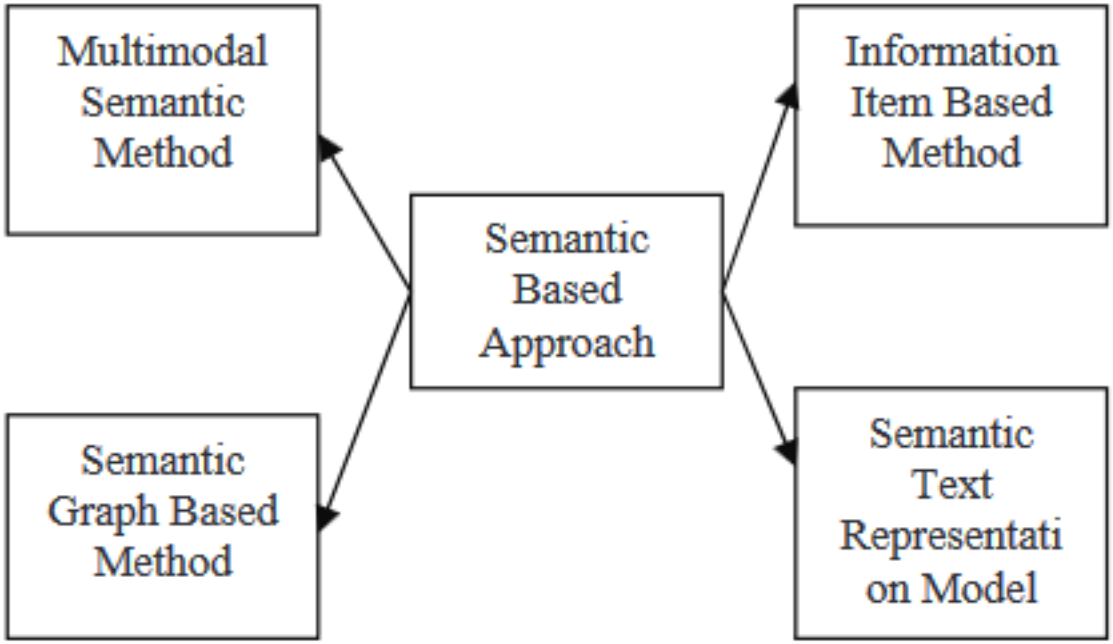
STRUCTURE-BASED APPROACH

- **Lead and body phrase method**
  - This approach involves rewriting the lead sentence by performing operations on phrases (insertion and substitution) with the same syntactic head chunk in the lead and body of the sentence.
  - The insertion process entails choosing an insertion point, checking for redundancy, and checking the discourse for internal coherence to ensure coherency and elimination of redundancy.
  - The substitution step provides increased information by substituting the body phrase in the lead chunk.

STRUCTURE-BASED APPROACH

- **Rule-based method**
  - In this technique, the documents to be summarized are depicted in terms of classes and listing of aspects.
  - The content choice module selects the most effective candidate among those generated by data extraction rules to answer one or many aspects of a category.
  - Finally, generation patterns are used for the generation of outline sentences.

SEMANTIC-BASED APPROACH



SEMANTIC-BASED APPROACH

- **Multimodal semantic model**
  - In this method, a linguistics model that captures concepts and relationships between ideas is created to describe the contents of multimodal documents such as text and images.
  - The key ideas are rated using several criteria, and the selected concepts are then expressed as sentences to form a summary.
- **Information item-based method**
  - In this approach, rather than using sentences from the supply documents, an abstract representation of those documents is used to generate the summary's content.
  - The abstract illustration is an Information item, the smallest part of coherent information in a text.

SEMANTIC-BASED APPROACH

- **Semantic Graph Model**
  - This technique aims to summarize a document by building a rich semantic graph (RSG) for the initial document, then reducing the created linguistics graph and generating the final abstractive outline from the reduced linguistics graph.
- **Semantic Text Representation Model**
  - This technique analyzes input text using words' semantics rather than the syntax/Structure of text.

Sentence Simplification

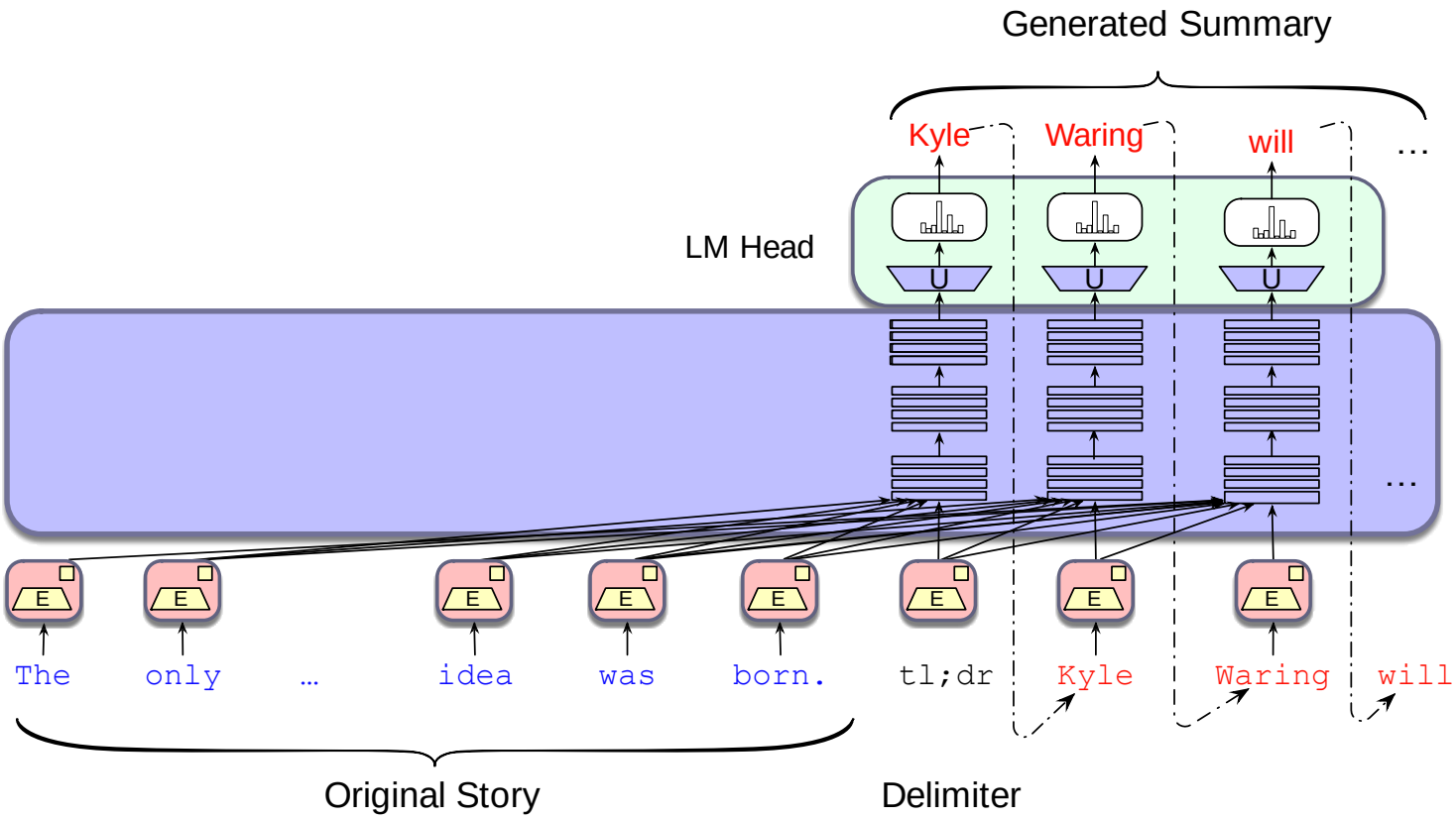
Once a set of sentences has been extracted and ordered, the final step in single-document summarization is **sentence realization**.

- One component of sentence realization is **sentence compression** or **sentence simplification**. The following examples, taken from a human summary, show that the human summarizer chose to eliminate some of the adjective modifiers and subordinate clauses when expressing the extracted sentence in the summary:
- **Original sentence:** When it arrives sometime new year in new TV sets, the V-chip will give parents a new and potentially revolutionary device to block out programs they don't want their children to see.
- **Simplified sentence by humans:** The V-chip will give parents a device to block out programs they don't want their children to see.

Sentence Simplification

- The simplest algorithms for sentence simplification use rules to select parts of the sentence to prune or keep, often by running a parser or partial parser over the sentences.
- More sophisticated models of sentence compression are based on supervised machine learning, in which a parallel corpus of documents together with their human summaries is used to compute the probability that particular words or parse nodes will be pruned

LLMs for summarization (using tl;dr)



Evaluation metrics

- BLEU
- ROUGE

BLEU (bilingual evaluation understudy)

- **BLEU** is an evaluation metric commonly used in machine translation.
- It measures similarity between ground truth and model output for a sequence of *n* words, known as n-grams.
- In text summarization, BLEU measures how often, and to what extent, n-grams in an automatic summary overlap with those in a human-generated summary, accounting for erroneous word repetitions in the former.
- It then uses these precision scores for individual n-grams to calculate an overall text precision, known as the geometric mean precision.
- This final value is between 0 and 1, the latter indicating perfect alignment between the machine and human generated text

ROUGE (recall-oriented understudy for gisting evaluation)

- **ROUGE** is derived from BLEU specifically for evaluating summarization tasks.
- Like BLEU, it compares machine summaries to human-generated summaries using n-grams.
- But while BLEU measures machine precision, ROUGE measures machine recall.
- In other words, ROUGE computes the accuracy of an automatic summary according to the number of n-grams from the human-generated summarization found in the automatic summary.
- The ROUGE score, like BLEU, is any value between 0 and 1, the latter indicating perfect alignment between the machine and human generated text summaries.

Applications of Text Summarization

- Research
- Customer feedback
- Corporate meetings
- Internal Documents
- Financial research

Sequence Labeling for Part of Speech and Named Entities

Named Entity Recognition (NER)

Named Entities

- **Named entity**, in its core usage, means anything that can be referred to with a proper name. Most common 4 tags:
  - PER (Person): “Marie Curie”
  - LOC (Location): “New York City”
  - ORG (Organization): “Stanford University”
  - GPE (Geo-Political Entity): "Boulder, Colorado"
- Often multi-word phrases
- But the term is also extended to things that aren't entities:
  - dates, times, prices

Named Entity tagging

The task of named entity recognition (NER):

- find spans of text that constitute proper names
- tag the type of the entity.

NER output

Citing high fuel prices, [ORG **United Airlines**] said [TIME **Friday**] it has increased fares by [MONEY **$6**] per round trip on flights to some cities also served by lower-cost carriers. [ORG **American Airlines**], a unit of [ORG **AMR Corp.**], immediately matched the move, spokesman [PER **Tim Wagner**] said. [ORG **United**], a unit of [ORG **UAL Corp.**], said the increase took effect [TIME **Thursday**] and applies to most routes where it competes against discount carriers, such as [LOC **Chicago**] to [LOC **Dallas**] and [LOC **Denver**] to [LOC **San Francisco**].

Why NER?

Sentiment analysis: consumer's sentiment toward a particular company or person?

Question Answering: answer questions about an entity?

Information Extraction: Extracting facts about entities from text.

Why NER is hard

1) Segmentation
   - In POS tagging, no segmentation problem since each word gets one tag.
   - In NER we have to find and segment the entities!
2) Type ambiguity

[PER Washington] was born into slavery on the farm of James Burroughs.
[ORG Washington] went up 2 games to 1 in the four-game series.
Blair arrived in [LOC Washington] for what may well be his last state visit.
In June, [GPE Washington] passed a primary seatbelt law.

BIO Tagging

How can we turn this structured problem into a sequence problem like POS tagging, with one label per word?

[PER Jane Villanueva] of [ORG United] , a unit of [ORG United Airlines Holding] , said the fare applies to the [LOC Chicago] route.

BIO Tagging

[PER Jane Villanueva] of [ORG United] , a unit of [ORG United Airlines Holding] , said the fare applies to the [LOC Chicago] route.

Words	BIO Label
Jane	B-PER
Villanueva	I-PER
of	O
United	B-ORG
Airlines	I-ORG
Holding	I-ORG
discussed	O
the	O
Chicago	B-LOC
route	O
.	O

Now we have one tag per token!!!

BIO Tagging

B: token that *begins* a span

I: tokens *inside* a span

O: tokens outside of any span

# of tags (where n is #entity types):

1 O tag,

n B tags,

n I tags

total of *2n+1*

Words	BIO Label
Jane	B-PER
Villanueva	I-PER
of	O
United	B-ORG
Airlines	I-ORG
Holding	I-ORG
discussed	O
the	O
Chicago	B-LOC
route	O
.	O

BIO Tagging variants: IO and BIOES

[PER Jane Villanueva] of [ORG United] , a unit of [ORG United Airlines Holding] , said the fare applies to the [LOC Chicago] route.

Words	IO Label	BIO Label	BIOES Label
Jane	I-PER	B-PER	B-PER
Villanueva	I-PER	I-PER	E-PER
of	O	O	O
United	I-ORG	B-ORG	B-ORG
Airlines	I-ORG	I-ORG	I-ORG
Holding	I-ORG	I-ORG	E-ORG
discussed	O	O	O
the	O	O	O
Chicago	I-LOC	B-LOC	S-LOC
route	O	O	O
.	O	O	O

Standard algorithms for NER

Supervised Machine Learning given a human-labeled training set of text annotated with tags

- Hidden Markov Models
- Conditional Random Fields (CRF)/ Maximum Entropy Markov Models (MEMM)
- Neural sequence models (RNNs or Transformers)
- Large Language Models (like BERT), finetuned

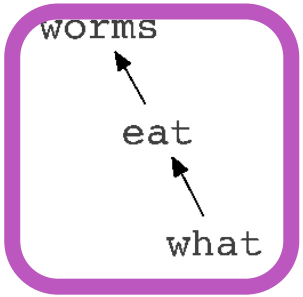
Question Answering

Question Answering

One of the oldest NLP tasks (punched card systems in 1961)

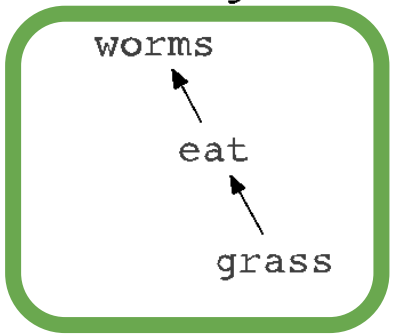
964. Indexing and
English Questions.
, 196-204

Question: What do worms eat?

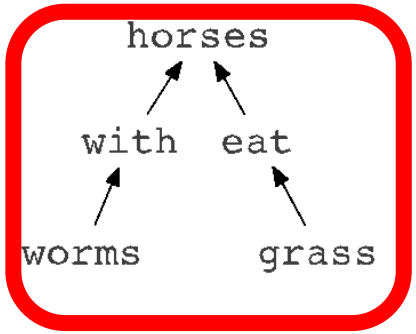


Potential Answers:

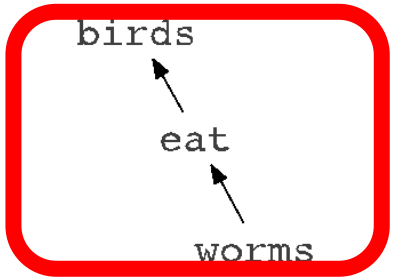
Worms eat grass



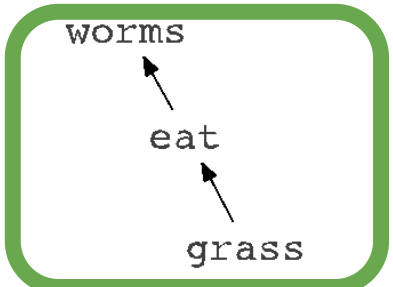
Horses with worms eat grass



Birds eat worms



Grass is eaten by worms



Question answering in TREC

- annual competition
- TREC-1 - 1992
- many QA approaches start there
- factoid and complex questions

Question Answering: IBM's Watson

- Won Jeopardy on February 16, 2011!

WILLIAM WILKINSON'S
"AN ACCOUNT OF THE PRINCIPALITIES
OF
WALLACHIA AND MOLDOVIA"
INSPIRED THIS AUTHOR'S
MOST FAMOUS NOVEL

Bram Stoker

Question Answering: IBM's Watson

Rhyme time category

It's where Pele stores his ball.

soccer locker

Before and After Goes to the Movies

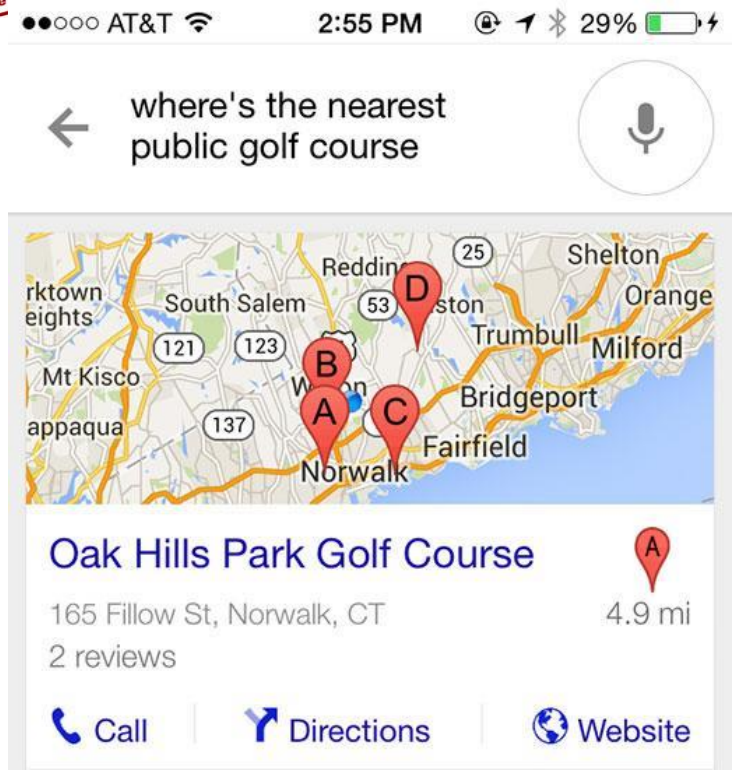
Film of a typical day in the life of the Beatles, which includes running from bloodthirsty zombie fans in a Romero classics.

A Hard Day's Night of the Living Dead

Apple's Siri


AT&T
22:26
65 %
“ Do I need an umbrella tomorrow in San Francisco ”
Yes, San Francisco should get rain tomorrow:
Weekly Forecast
TUES WED THU FRI SAT SUN
14° 16° 17° 17° 18° 18°
9° 7° 6° 6° 7° 8°

Google now


AT&T
2:55 PM
29%
where's the nearest public golf course
Shelton
Orange
South Salem
Trumbull
Milford
Mt Kisco
Bridgeport
Fairfield
Norwalk
Oak Hills Park Golf Course
165 Fillow St, Norwalk, CT
4.9 mi
2 reviews
Call
Directions
Website


AT&T
2:45 PM
19%
what's the five day weathe…
Wilton, CT 06897
Wed, Partly Cloudy
63 °F | °C
Precip: 10%
Humidity: 61%
Wind: 9 mph
4 PM
9 PM
2 AM
7 AM
12 PM
WED
THU
FRI
SAT
SUN
MON
63° 34°
48° 30°
43° 27°
43° 27°
41° 32°
41° 23°

WolframAlpha computational knowledge engine

how many calories are in two slices of banana cream pie?

Examples Random

Assuming any type of pie, banana cream | Use pie, banana cream, prepared from recipe or pie, banana cream, no-bake type, prepared from mix instead

Input interpretation:

pie	amount	2 slices	total calories
	type	banana cream	

Average result: Show details

702 Cal (dietary Calories)

Types of Questions in Modern Systems

- Factoid questions
  - *Who wrote "The Universal Declaration of Human Rights"?*
  - *How many calories are there in two slices of apple pie?*
  - *What is the average age of the onset of autism?*
  - *Where is Apple Computer based?*
- Complex (narrative) questions:
  - *In children with an acute febrile illness, what is the efficacy of acetaminophen in reducing fever?*
  - *What do scholars think about Jefferson's position on dealing with pirates?*

Commercial systems: mainly factoid questions

Where is the Louvre Museum located?	In Paris, France
What's the abbreviation for limited partnership?	L.P.
What are the names of Odin's ravens?	Huginn and Muninn
What currency is used in China?	The yuan
What kind of nuts are used in marzipan?	almonds
What instrument does Max Roach play?	drums

Paradigms for QA

- IR-based approaches
  - TREC; IBM Watson; Google
- Knowledge-based and Hybrid approaches
  - IBM Watson; Apple Siri; Wolfram Alpha; True Knowledge Evi

Many questions can already be answered by web search

Google What are the names of Odin's ravens?

Search About 214,000 results (0.38 seconds)

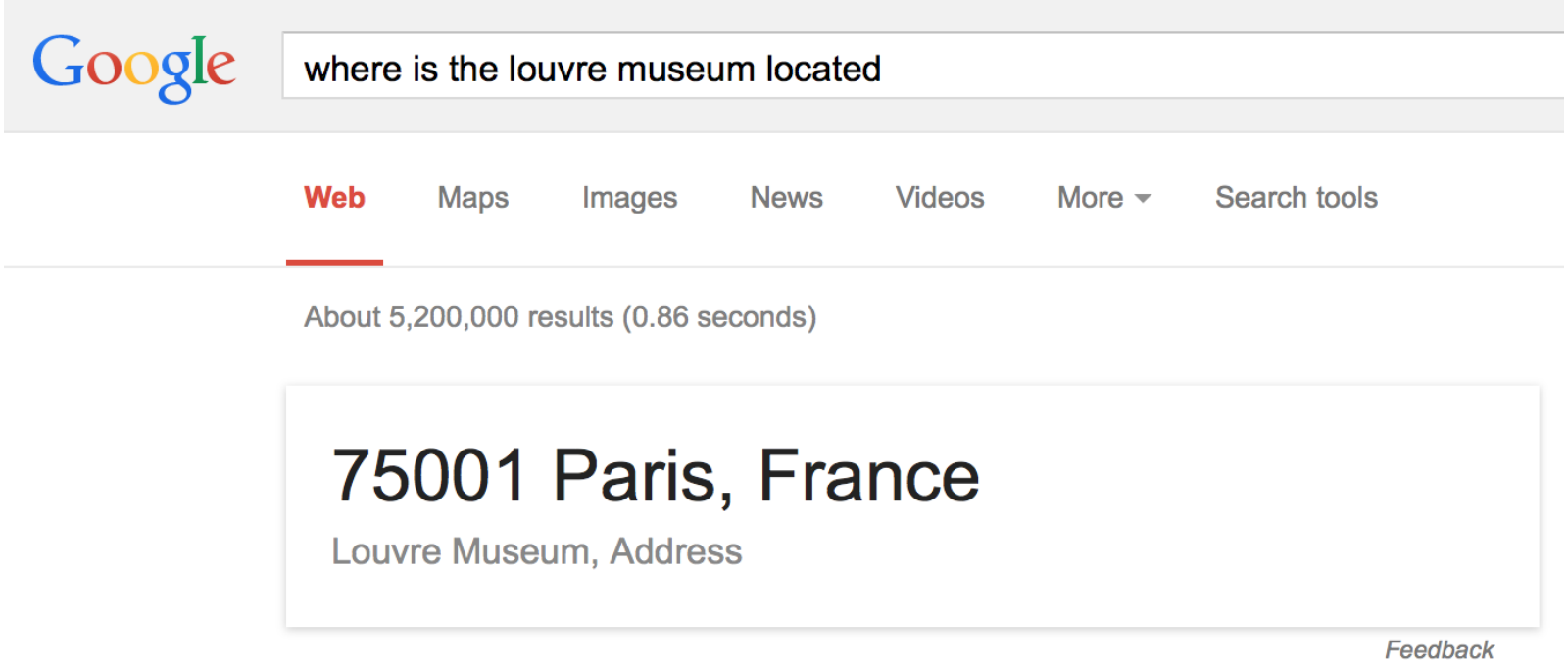
Everything

Images

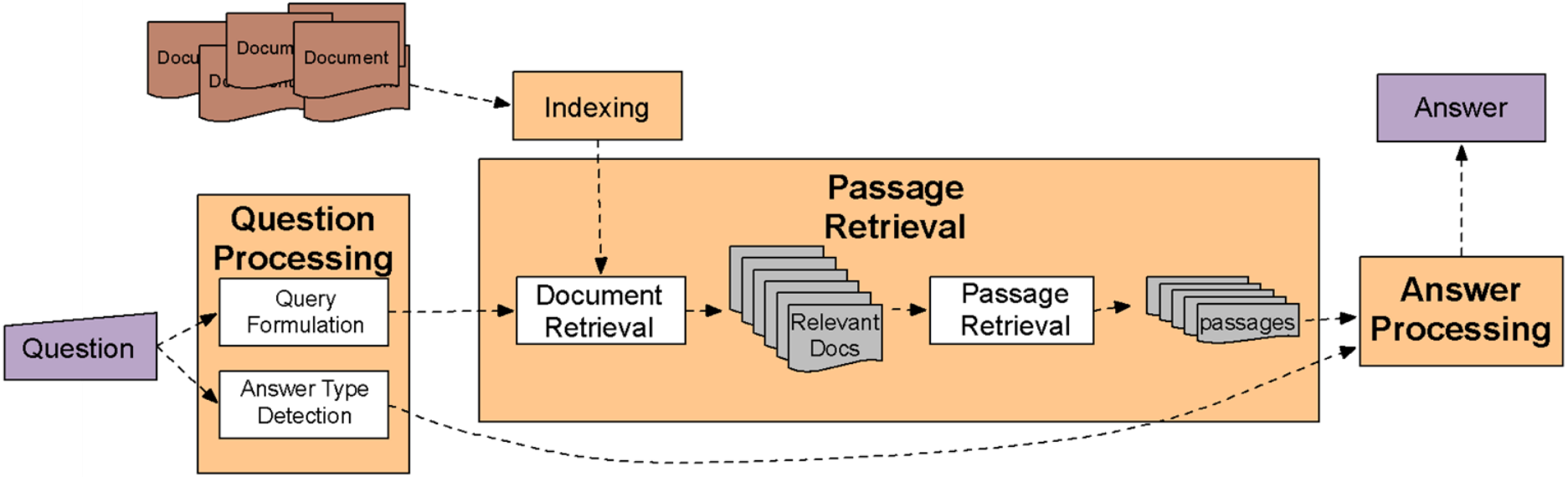
Maps

Huginn and Muninn - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Huginn_and_Muninn
The **names** of the **ravens** are sometimes modernly anglicized as Hugin and Munin. In the Poetic Edda, a disguised **Odin** expresses that he fears that they may ...
Attestations - Archaeological record - Theories - See also

IR-based Question Answering



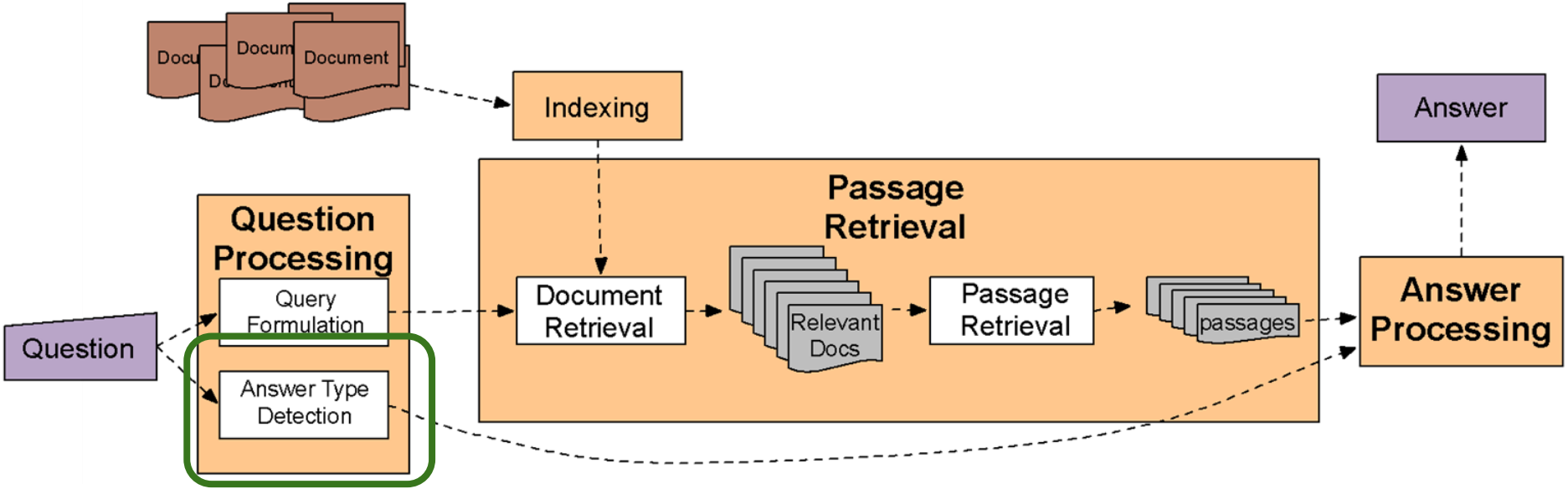
IR-based Factoid QA



IR-based Factoid QA

- QUESTION PROCESSING
  - Detect question type, answer type, focus, relations
  - Formulate queries to send to a search engine
- PASSAGE RETRIEVAL
  - Retrieve ranked documents
  - Break into suitable passages and rerank
- ANSWER PROCESSING
  - Extract candidate answers
  - Rank candidates
    - using evidence from the text and external sources

IR-based Factoid QA



Question Processing
Things to extract from the question

- Answer Type Detection
  - Decide the **named entity type** (person, place) of the answer
- Query Formulation
  - Choose **query keywords** for the IR system
- Question Type classification
  - Is this a definition question, a math question, a list question?
- Focus Detection
  - Find the question words that are replaced by the answer
- Relation Extraction
  - Find relations between entities in the question

Question Processing

They're the two states you could be reentering if you're crossing Florida's northern border

- Answer Type: US state
- Query: two states, border, Florida, north
- Focus: the two states
- Relations: borders(Florida, ?x, north)

Answer Type Detection: Named Entities

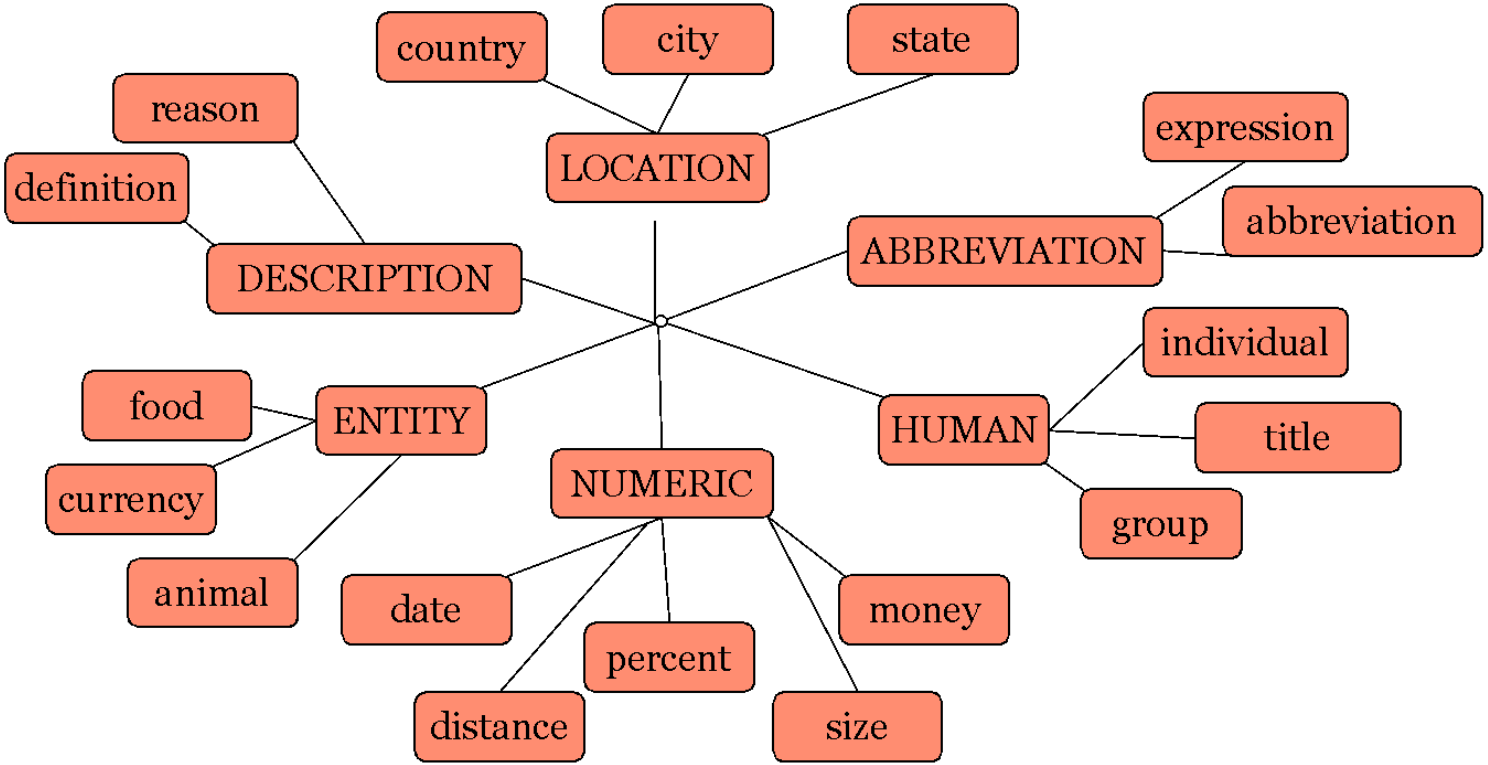
- *Who founded Virgin Airlines?*
  - PERSON
- *What Canadian city has the largest population?*
  - *CITY.*

Answer Type Taxonomy

Xin Li, Dan Roth. 2002. Learning Question Classifiers. COLING'02

- 6 coarse classes
  - ABBEVIATION, ENTITY, DESCRIPTION, HUMAN, LOCATION, NUMERIC
- 50 finer classes
  - LOCATION: city, country, mountain…
  - HUMAN: group, individual, title, description
  - ENTITY: animal, body, color, currency…

Part of Li & Roth's Answer Type Taxonomy



Answer Types

ENTITY	
animal	What are the names of Odin's ravens?
body	What part of your body contains the corpus callosum?
color	What colors make up a rainbow ?
creative	In what book can I find the story of Aladdin?
currency	What currency is used in China?
disease/medicine	What does Salk vaccine prevent?
event	What war involved the battle of Chapultepec?

Answer Type Detection

- Hand-written rules
- Machine Learning
- Hybrids

Answer Type Detection

- Regular expression-based rules can get some cases:
  - Who {is|was|are|were} PERSON
  - PERSON (YEAR – YEAR)
- Other rules use the **question headword**:

  (the headword of the first noun phrase after the wh-word)

  - Which **city** in China has the largest number of foreign financial companies?
  - What is the state **flower** of California?

Answer Type Detection

- Most often, we treat the problem as machine learning classification
  - **Define** a taxonomy of question types
  - **Annotate** training data for each question type
  - **Train** classifiers for each question class using a rich set of features.
    - features include those hand-written rules!

Features for Answer Type Detection

- Question words and phrases
- Part-of-speech tags
- Parse features (headwords)
- Named Entities
- Semantically related words

Answer Type Detection

Xin Li, Dan Roth. 2002. Learning Question Classifiers. COLING'02

Goal: categorize questions into different semantic classes that impose constraints on potential answers.

Who was the first woman killed in Vietnam war? - *PERSON*

What is a prism? - *DEFINITION*

Why is the sun yellow? - *REASON*

Answer Type Detectior

Xin Li, Dan Roth. 2002. Learning Question Classifiers. COLING'02

Existing approaches perform coarse classification (20 classes).

Li and Roth define a finer taxonomy of answer types.

Class	#	Class	#
ABBREV.	9	description	7
abb	1	manner	2
exp	8	reason	6
ENTITY	94	**HUMAN**	65
animal	16	group	6
body	2	individual	55
color	10	title	1
creative	0	description	3
currency	6	**LOCATION**	81
dis.med.	2	city	18
event	2	country	3
food	4	mountain	3
instrument	1	other	50
lang	2	state	7
letter	0	**NUMERIC**	113
other	12	code	0
plant	5	count	9
product	4	date	47
religion	0	distance	16
sport	1	money	3
substance	15	order	0
symbol	0	other	12
technique	1	period	8
term	7	percent	3
vehicle	4	speed	6
word	0	temp	5
DESCRIPTION	138	size	0
definition	123	weight	4

Answer Type Detection: Ambiguity

Xin Li, Dan Roth. 2002. Learning Question Classifiers.
COLING'02

What is bipolar disorder? - *definition* or *desease_medicine*

What do bats eat? - *food*, *plant*, or *animal*

What is PH scale? - *numeric_value* or *definition*

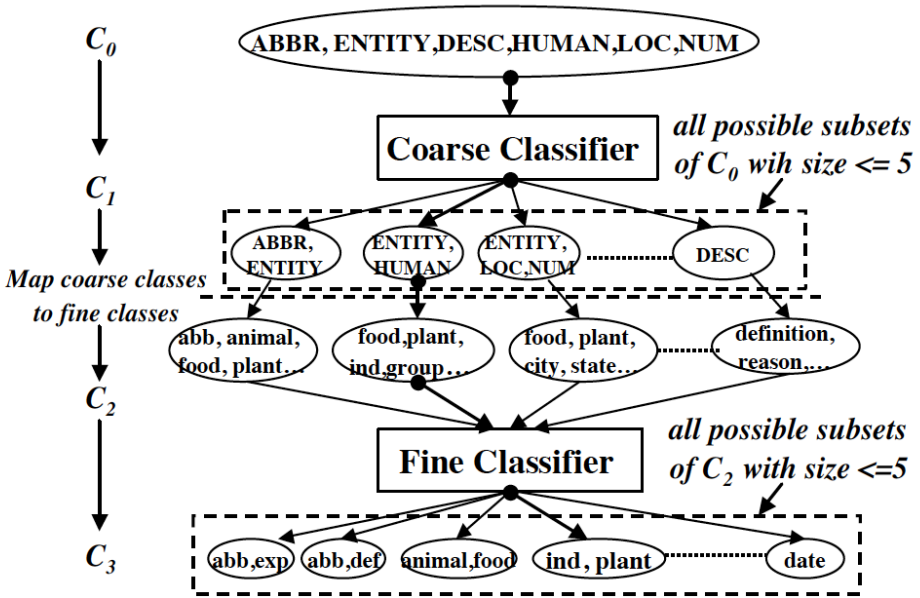
One question is allowed to have several class labels

Answer Type Detection: Building a classifier

Xin Li, Dan Roth. 2002. Learning Question Classifiers.
COLING'02

Two classifiers:

- Coarse classifier
- Fine classifier



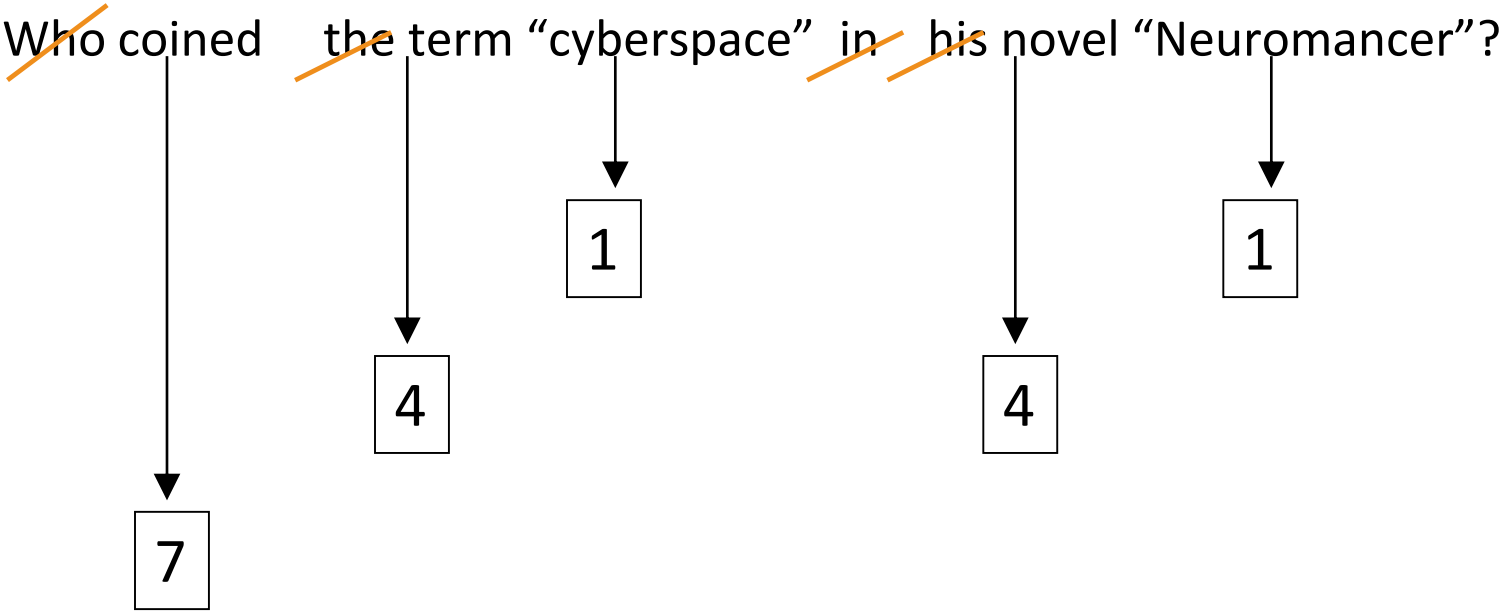
Keyword Selection Algorithm

Dan Moldovan, Sanda Harabagiu, Marius Paca, Rada Mihalcea, Richard Goodrum, Roxana Girju and Vasile Rus. 1999. Proceedings of TREC-8.

1. Select all non-stop words in quotations
2. Select all NNP words in recognized named entities
3. Select all complex nominals with their adjectival modifiers
4. Select all other complex nominals
5. Select all nouns with their adjectival modifiers
6. Select all other nouns
7. Select all verbs
8. Select all adverbs
9. Select the QFW word (skipped in all previous steps)
10. Select all other words

Choosing keywords from the query

Slide from Mihai Surdeanu



cyberspace/1 Neuromancer/1 term/4 novel/4 coined/7

Passage Retrieval

- Step 1: IR engine retrieves documents using query terms
- Step 2: Segment the documents into shorter units
  - something like paragraphs
- Step 3: Passage ranking
  - Use answer type to help rerank passages

Features for Passage Ranking

Either in rule-based classifiers or with supervised machine learning

- Number of Named Entities of the right type in passage
- Number of query words in passage
- Number of question N-grams also in passage
- Proximity of query keywords to each other in passage
- Longest sequence of question words
- Rank of the document containing passage

Answer Extraction

- Run an answer-type named-entity tagger on the passages
  - Each answer type requires a named-entity tagger that detects it
  - If answer type is CITY, tagger has to tag CITY
    - Can be full NER, simple regular expressions, or hybrid
- Return the string with the right type:
  - Who is the prime minister of India (PERSON)

Manmohan Singh, Prime Minister of India, had told left leaders that the deal would not be renegotiated.

  - How tall is Mt. Everest? (LENGTH)

The official height of Mount Everest is **29035 feet**

Ranking Candidate Answers

- But what if there are multiple candidate answers!

Q: Who was Queen Victoria's second son?

- Answer Type: **Person**

- Passage:

The Marie biscuit is named after Marie Alexandrovna, the daughter of Czar Alexander II of Russia and wife of Alfred, the second son of Queen Victoria and Prince Albert

Ranking Candidate Answers

- But what if there are multiple candidate answers!

Q: Who was Queen Victoria's second son?

- Answer Type: **Person**
- Passage:

The Marie biscuit is named after **Marie Alexandrovna**, the daughter of **Czar Alexander II of Russia** and wife of **Alfred**, the second son of **Queen Victoria** and **Prince Albert**

Use machine learning: Features for ranking candidate answers

Answer type match: Candidate contains a phrase with the correct answer type.

Pattern match: Regular expression pattern matches the candidate.

Question keywords: # of question keywords in the candidate.

Keyword distance: Distance in words between the candidate and query keywords

Novelty factor: A word in the candidate is not in the query.

Apposition features: The candidate is an appositive to question terms

Punctuation location: The candidate is immediately followed by a comma, period, quotation marks, semicolon, or exclamation mark.

Sequences of question terms: The length of the longest sequence of question terms that occurs in the candidate answer.

Knowledge-based approaches

- Build a semantic representation of the query
  - Times, dates, locations, entities, numeric quantities
- Map from this semantics to query structured data or resources
  - Geospatial databases
  - Ontologies (Wikipedia infoboxes, dbPedia, WordNet, Yago)
  - Restaurant review sources and reservation services
  - Scientific databases

Relation Extraction

- Answers: Databases of Relations
  - born-in(“Emma Goldman”, “June 27 1869”)
  - author-of(“Cao Xue Qin”, “Dream of the Red Chamber”)
  - Draw from Wikipedia infoboxes, DBpedia, FreeBase, etc.
- Questions: Extracting Relations in Questions

  Whose granddaughter starred in E.T.?

```
(acted-in ?x “E.T.”)
   (granddaughter-of ?x ?y)
```

Temporal Reasoning

- Relation databases
  - (and obituaries, biographical dictionaries, etc.)
- IBM Watson

  ”In 1594 he took a job as a tax collector in Andalusia”

  Candidates:
  - Thoreau is a bad answer (born in 1817)
  - Cervantes is possible (was alive in 1594)

Geospatial knowledge (containment, directionality, borders)

- Beijing is a good answer for "Asian city"
- California is "southwest of Montana"
- geonames.org: