Natural Language Processing

Course: CSE 537
Instructor: Professor Anita Wasilewska

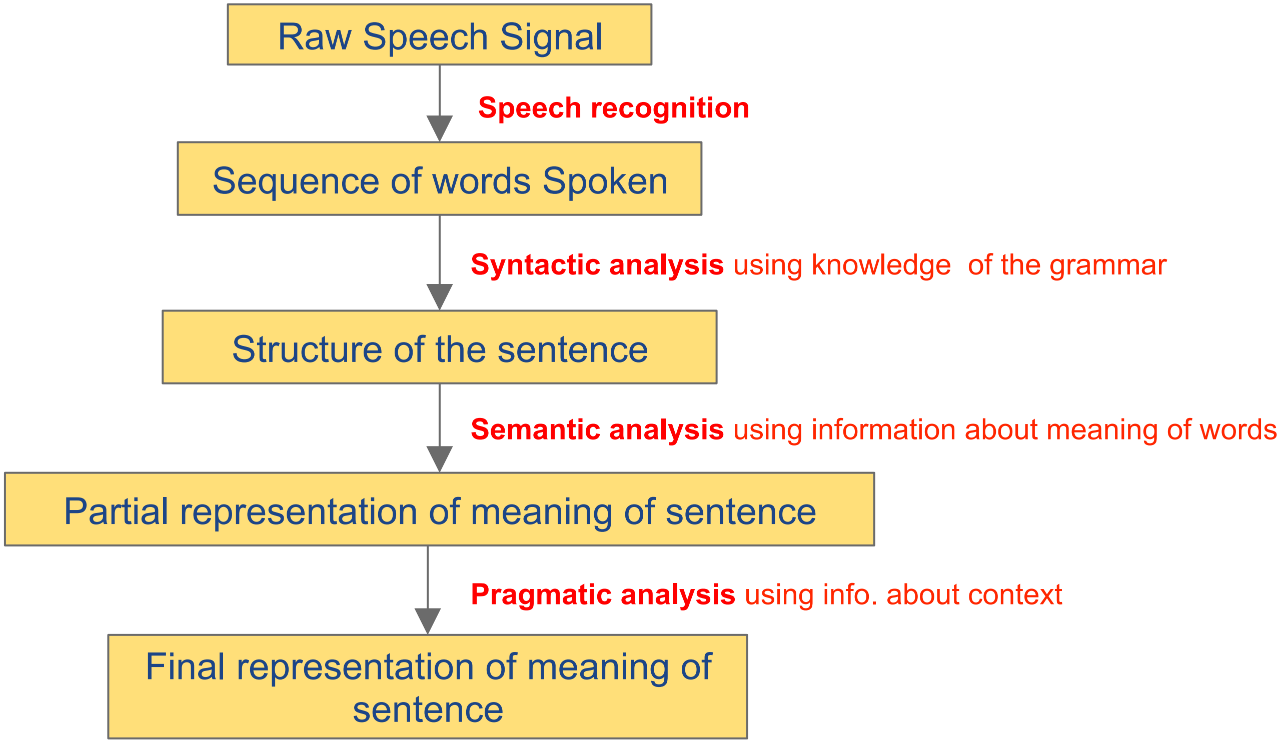
Group 11:
Garima Gehlot (110167347)
Mansi Agarwal (110395542)
Rahul Venkataraman (110368788)

What is Natural Language Processing?

- **Natural Language Processing (NLP)** is the branch of computer science focused on developing systems that allow computers to communicate with people using everyday language.
- Natural language processing is a subfield of Artificial Intelligence and linguistic, devoted to make computers “understand” statements written in human languages.
- Also known as **Computational Linguistics**.

Abhimanyu Chopra, Abhinav Prashar, and Chandresh Sain. *Natural language processing. International Journal of Technology Enhancements and Emerging Engineering Research*, 1(4):131–134, 2013.

Natural Language Understanding



Raw Speech Signal

Speech recognition

Sequence of words Spoken

Syntactic analysis using knowledge of the grammar

Structure of the sentence

Semantic analysis using information about meaning of words

Partial representation of meaning of sentence

Pragmatic analysis using info. about context

Final representation of meaning of sentence

Reference: Prof. Carolina Ruiz NLP slides (web.cs.wpi.edu/~cs534/f07/LectureNotes/Slides/nat_lang_processing.ppt)

Speech Recognition

- Spoken language is recognized and transformed into text as in dictation systems, into commands as in robot control systems, or into some other internal representation.
- Get sound input using microphone, sample and convert to digital data.
- Segment the sounds to recognizable sounds.

Syntactic Analysis

Rules of syntax (grammar) specify the possible organization of words in sentences and allows us to determine sentence’s structure(s).

Parsing

Checks that the sentence is correct according to the grammar and if so returns a **parse tree** representing the structure of the sentence.

Grammar

Sentence to verb_phrase and noun_phrase.

Identify verb and noun phrases in the sentence.

Decompose it further into proper_nouns, verbs, determiner.

Reference: Prof. Carolina Ruiz NLP slides (web.cs.wpi.edu/~cs534/f07/LectureNotes/Slides/nat_lang_processing.ppt)

Syntactic Analysis - Grammar

E.g. John drank the soup.

Sentence (S) **------->** *verb_phrase (VP), noun_phrase (NP)*

NP --------> *Noun (N)*

NP --------> *Article, Noun*

VP --------> *verb (V), NP*

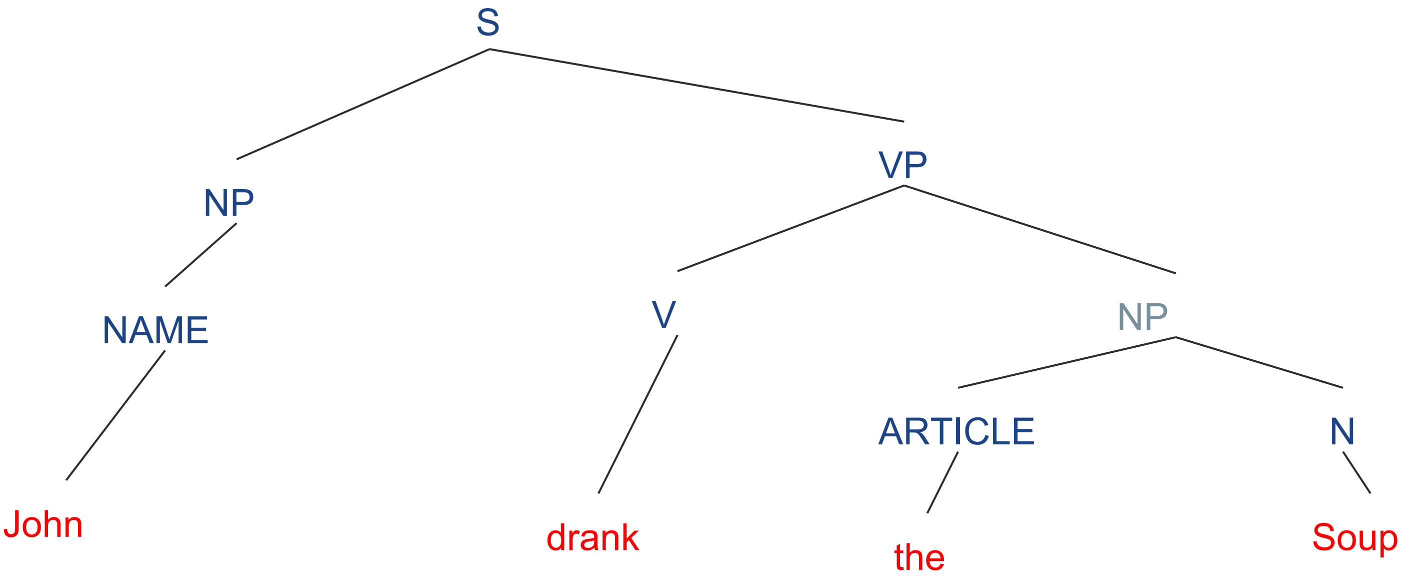
N ----------> *[John]*

N ------> *[soup]*

V --------> *[drank]*

Article --------> *[the]*

Syntactic Analysis - Parsing



Semantic Analysis

- Generates (partial) meaning/representation of the sentence from its syntactic structure.
- Lexical Semantics
    - Meanings of component words
    - Word sense disambiguation
- Compositional Semantics
    - How words combine to form larger meanings
- Roughly:
  Semantic Analysis ≈ Understanding Language

Book: Digital Content Annotation and Transcoding *by Katashi Nagao, Artech House, Inc. Norwood, MA, USA ©2003.*

Semantic Analysis (contd.)

Lexical Processing

The first step in any semantic processing system is to look up the individual words in a dictionary (or Lexicon) and extract their meanings.

Many words have several meanings and it may not be possible to choose the correct meaning just by looking at the word.

The process of determining the correct meaning of an individual word is called Word Sense disambiguation or lexical disambiguation.

It is done by associating, with each word in lexicon, information about the contexts in which each of the word may appear.

Book: Digital Content Annotation and Transcoding *by Katashi Nagao, Artech House, Inc. Norwood, MA, USA ©2003.*

Approaches to Semantic Analysis

→ *Predicate logic*

- For Example, the sentence
  ”a restaurant that serves Chinese food near TUT”
  corresponds to the meaning representation:

$$\exists x\ Restaurant(x) \land Serves(x, ChineseFood) \land Near(LocationOf(x), LocationOf(TUT))$$

- Logical propositions enable inference
- scalability problem (large vocabulary/unrestricted domain)

Approaches to Semantic Analysis

➔ Statistical Approach

- A database
- Learn an alignment: words and phrases that correspond to each other
- Learn word order in the target language (probabilities of target word strings)
- Translate by matching source fragments against a database of real world examples, identifying the corresponding translation fragments, and then recombining these to give the target text.

Approaches to Semantic Analysis

➔ Domain Knowledge Driven Analysis

- Expect certain slots of information to be filled in.
- Restricting to a certain domain allows use of specific patterns, rules, expectations etc.
    - Customer at a restaurant
    - Buying train tickets etc.

Approaches to Semantic Analysis

➔ Information Retrieval

- Google solves a certain part of the problem in a statistical way: answers to ”trivial” kind of questions can be located using a web search engine

- assumes a database (Internet) and a clever page ranking system.

Information Extraction (IE)

Information Extraction is used to process a body of text so that it can be entered into a relational database or analyzed using data mining. IE systems generally focus on a specific domain or topic,searching only for information that is relevant to a user's interests.

Input: Body of texts.

Output: a closely defined data format that is suitable for a database or data mining application.

The rigid format for the final result means that only a fraction of the data is relevant.

Information Extraction Literature

Entity *- an object of interest such as a person or an organization. In object oriented programming terminology an entity would be an object, like a person.*

Attribute *A property of an entity, such as the entity name, an alias or the entity type.An example of entities might be people. Attributes are the details of the person and their relations to other people.*

Dictionary *A set of case frames.*

Book: Natural Language Processing for Online Applications *by Peter Jackson and Isabelle Moulinier, John Benjamins Publishing Company, 2002.*

Evaluation of an IE system

A system's performance is measured on:

Recall (R) - what percentage of the correct answers did the system get

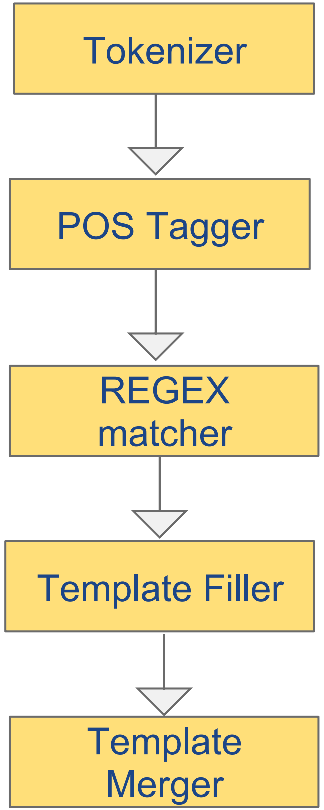
Precision (P) - what percentage of the system's answers were correct

F-score (F)- is a weighted harmonic mean between recall and precision computed by the following formula:

$$F_\beta = (1 + \beta^2) \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{\beta^2 \cdot \mathrm{precision} + \mathrm{recall}}$$

β is a parameter, encoding the relative importance of recall and precision.
If **β=1**, they are weighted equally.
If **β>1**, precision is more important.
if **β<1**, recall is more important.

Common Steps in Information Extraction



Book: Natural Language Processing for Online Applications *by Peter Jackson and Isabelle Moulinier, John Benjamins Publishing Company, 2002.*

Tokenizing (Text Splitting)

IE tokenizer is similar to a tokenizer or scanner in artificial language processing. However, the problem is more difficult with natural languages. Even simple constructs, like commas and periods add complexity.

For example, a tokenizer may have to recognize that the period in Mr. XYZ does not terminate the sentence.

Its purpose is to split the characters of the input document into disjoint meaningful parts.

Tokenizing is an easy task for programming languages, since they have a clearly defined grammar as well as defined output, but for natural languages it

Book: Natural Language Processing for Online Applications *by Peter Jackson and Isabelle Moulinier, John Benjamins Publishing Company, 2002.*

Part of Speech (POS)

Each word belongs to a word class. The word class of a word is known as part-of-speech (POS) of that word.

Most POS tags implicitly encode fine-grained specializations of eight basic parts of speech:

Noun, verb, pronoun, preposition, adjective, adverb, conjunction, article

These categories are based on morphological and distributional similarities (not semantic similarities).

Part of Speech is also known as: Word Classes, Lexical tags etc.

Part of Speech (contd.)

POS tag of a word describes the major and minor word classes of that word.

A POS tag of a word gives significant amount of information about that word and its neighbors.

For example, a personal pronoun (I, you, he, she) most likely will be followed by a verb.

Most of the words have single POS tag but some words have more than one. For example: book/*noun* or book/*verb*

I bought a book.

Please book *that flight.*

Book: A Resource-light Approach to Morpho-syntactic Tagging by Anna Feldman, Jirka Hana. Amsterdam: Rodopi (Language and computers: Studies in practical linguistics, volume 70), 2010.

Tag Sets

There are various tag sets to choose.

The choice of the tag set depends on the nature of the application.

Some of widely used POS tag sets are:

Penn Treebank has 45 tags

Brown Corpus has 87 tags

C7 has 146 tags

In tagged corpus, each word is associated with a tag from the used tag set.

Daniel Jurafsky and James H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*, (2nd Ed.), Prentice-Hall, 2008, 988 pages.

Part of Speech Tagging

Part of Speech tagging is simply assigning the correct part of the speech for each word in an input sentence.

Why Tagging is Hard?

Most words in English are unambiguous. They have only one tag.

But many of the most common words are ambiguous.

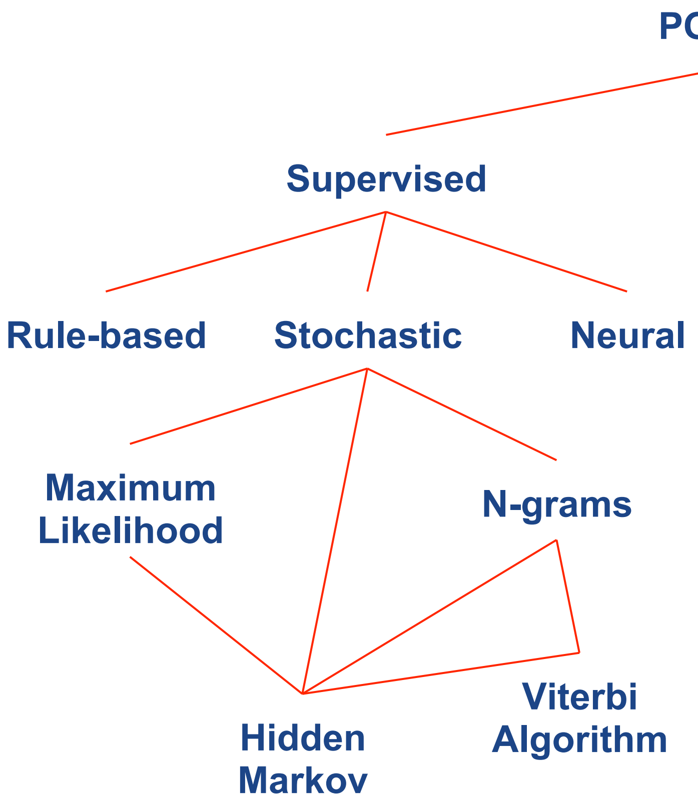
Can/verb Can/noun Can/auxillary

There are different algorithms for tagging:

Rule Based Tagging

Statistical Tagging (Stochastic Tagging)

POS tagging Approaches



Dinesh Kumar, Gurpreet Singh Josan. *September 2010.* Part of Speech Taggers for Morphologically Rich Indian Languages: A Survey. *International Journal of Computer Applications (0975 – 8887), Volume 6– No.5*

Rule Based Tagging

Typical rule based approaches use contextual information to assign tags to unknown or ambiguous words. These rules are often known as context frame rules. As an example, a context frame rule might say something like:

> *If an ambiguous/unknown word X is preceded by a determiner and followed by a noun, tag it as an adjective.*

In addition to contextual information, many taggers use morphological information to aid in the disambiguation process. One such rule might be:

> *if an ambiguous/unknown word ends in an -ing and is preceded by a verb, label it a verb (depending on your theory of grammar, of course).*

Some systems use contextual and morphological information by including rules pertaining to such factors as capitalisation and punctuation. (useful when language has significant properties in grammar).

Stochastic Tagging

Stochastic tagging means any model, which somehow incorporates frequency or probability.

The simplest stochastic taggers disambiguate words based solely on the probability that a word occurs with a particular tag.

An alternative to the word frequency approach is to calculate the probability of a given sequence of tags occurring.This is sometimes referred to as the ***n-gram*** approach.

- referring to the fact that the best tag for a given word is determined by the probability that it occurs with the '*n'* previous tags.

Evangelos Dermatas, George Kokkinakis. June 1995. Automatic stochastic tagging of natural language texts. *Computational Linguistics*. Volume 21 Issue 2, Pages 137-163

Stochastic Tagging (contd.)

The most common algorithm for implementing an n-gram approach is known as the **Viterbi Algorithm**, a search algorithm which avoids the polynomial expansion of a breadth first search by "pruning" the search tree at each level using the best N Maximum Likelihood Estimates (where n represents the number of tags of the following word).

Combination of the previous two approaches, using both tag sequence probabilities and word frequency measurements, is known as a **Hidden Markov Model.**

Evangelos Dermatas, George Kokkinakis. June 1995. Automatic stochastic tagging of natural language texts. *Computational Linguistics*. Volume 21 Issue 2, Pages 137-163

Hidden Markov Model (HMM)

Each hidden tag state produces a word in the sentence. Each word is:

- Uncorrelated with all the other words and their tags
- Probabilistic depending on the N previous tags only

Hidden Markov Model taggers may be implemented using the Viterbi algorithm, and are among the most efficient of the tagging methods discussed here.

HMM's cannot, however, be used in an automated tagging schema, since they rely critically upon the calculation of statistics on output sequences (tag-states).

Evangelos Dermatas, George Kokkinakis. June 1995. Automatic stochastic tagging of natural language texts. *Computational Linguistics*. Volume 21 Issue 2, Pages 137-163

Text Summarization

Identifies the most important points of a text and expresses them in a shorter human understandable language.

Process:

interpret the text

extract the relevant information

condense extracted information and create summary representation

present summary representation to reader in natural language

Summarization Applications

outlines or abstracts of any document, article,etc

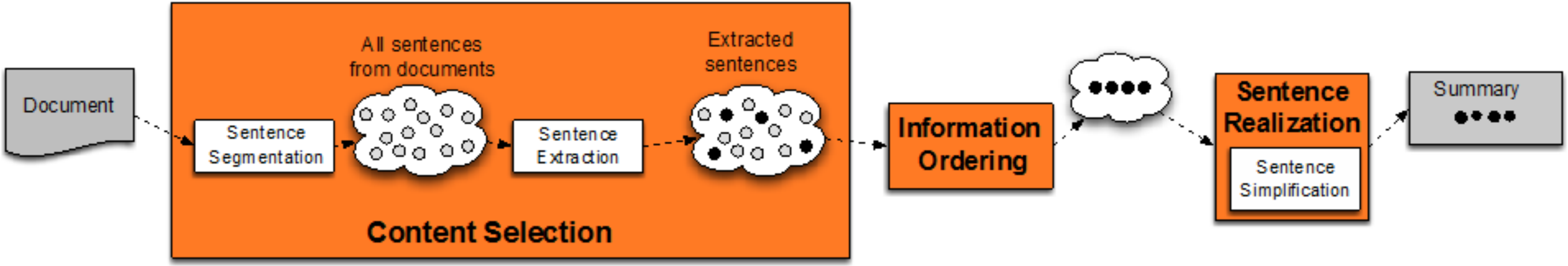
summaries of email threads

action items from a meeting

simplifying text by compressing sentences

Summarization: Three Stages

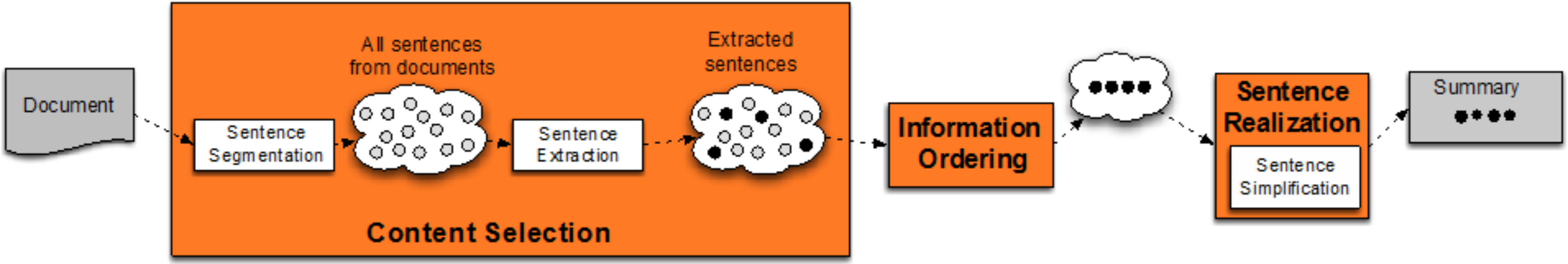
1. **Content selection**: choose sentences to extract from the document
2. **Information ordering**: choose an order to place them in the summary
3. **Sentence realization**: clean up the sentences



Reference: Prof. Chris Manning’s Stanford NLP Summarization Slides

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



Reference: Prof. Chris Manning’s Stanford NLP Summarization Slides

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 wi 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)

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

Reference: Prof. Chris Manning's Stanford NLP Summarization Slides

Topic signature-based Content Selection

choose words that are informative either

by log-likelihood ratio (LLR)

or by appearing in the query

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

Weigh a sentence (or window) by weight of its words:

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

Reference: Prof. Chris Manning’s Stanford NLP Summarization Slides

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)

Supervised Content Selection Issues

hard to get labeled training data

alignment difficult

performance not better than unsupervised algorithms

Conclusion: In practice

Unsupervised content selection is more common

Evaluate Summaries : ROUGE

ROUGE (Recall Oriented Understudy for Gisting Evaluation)

Intrinsic metric for automatically 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

ROGUE : How it works

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?

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

Reference: Prof. Chris Manning’s Stanford NLP Summarization Slides

ROGUE : Example

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.

$$\text{ROGUE-2} = \frac{3+3+6}{10+9+9} = 12/28 = .43$$

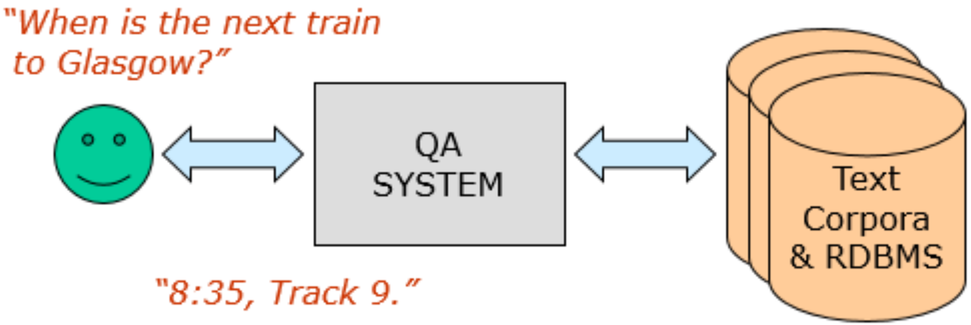
Reference: Prof. Chris Manning’s Stanford NLP Question Answering Slides

Question Answering:

One of the oldest NLP tasks.

Inputs: a question in English; a set of text and database resources.

Output: a set of possible answers drawn from the resources.



Picture Courtesy: CMU Natural Language Processing Slides: **15-381 Lecture, Spring 2003**

Why Question Answering?

QA engines attempt to let you ask your question the way you'd normally ask it.

More specific than short keyword queries.

Inexperienced search users.

People love asking questions!

Examples from Ask.com

how much should i weigh?

what does my name mean?

where can i find pictures of bikes?

who is the richest man in India?

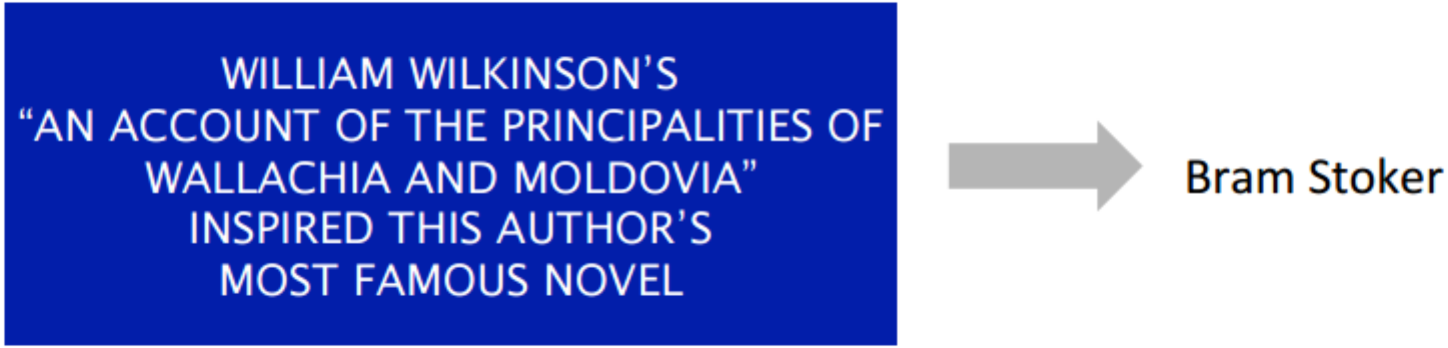
what is the meaning of life?

why is the sea blue?

can you drink milk after the expiration date?

Question Answering: IBM’s Watson

Won Jeopardy on February 16, 2011!

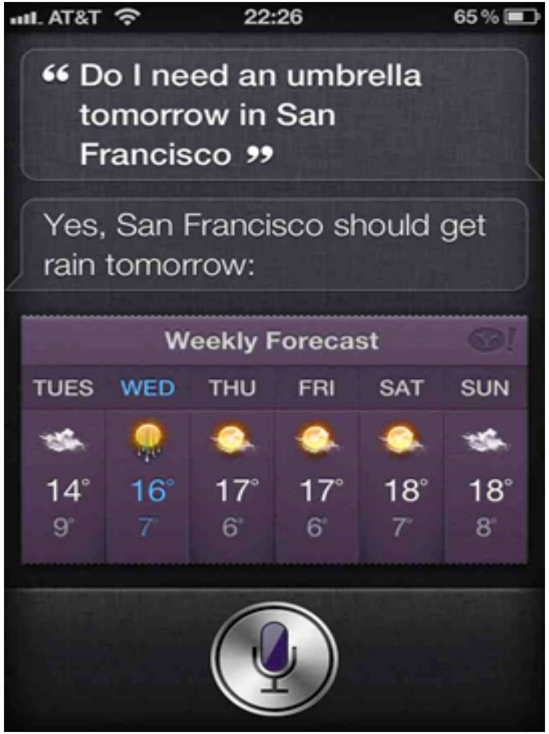


WILLIAM WILKINSON’S
“AN ACCOUNT OF THE PRINCIPALITIES OF
WALLACHIA AND MOLDOVIA”
INSPIRED THIS AUTHOR’S
MOST FAMOUS NOVEL

Bram Stoker

Reference: Prof. Chris Manning’s Stanford NLP Question Answering Slides

Question Answering: Apple’s Siri



“ 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°

Reference: Prof. Chris Manning’s Stanford NLP Question Answering Slides

Types of Questions in Modern Systems:

Factoid questions:

Who wrote “The Great Expectations”?

How many calories are there in two slices of Mango pie?

What is the average age when kids join school?

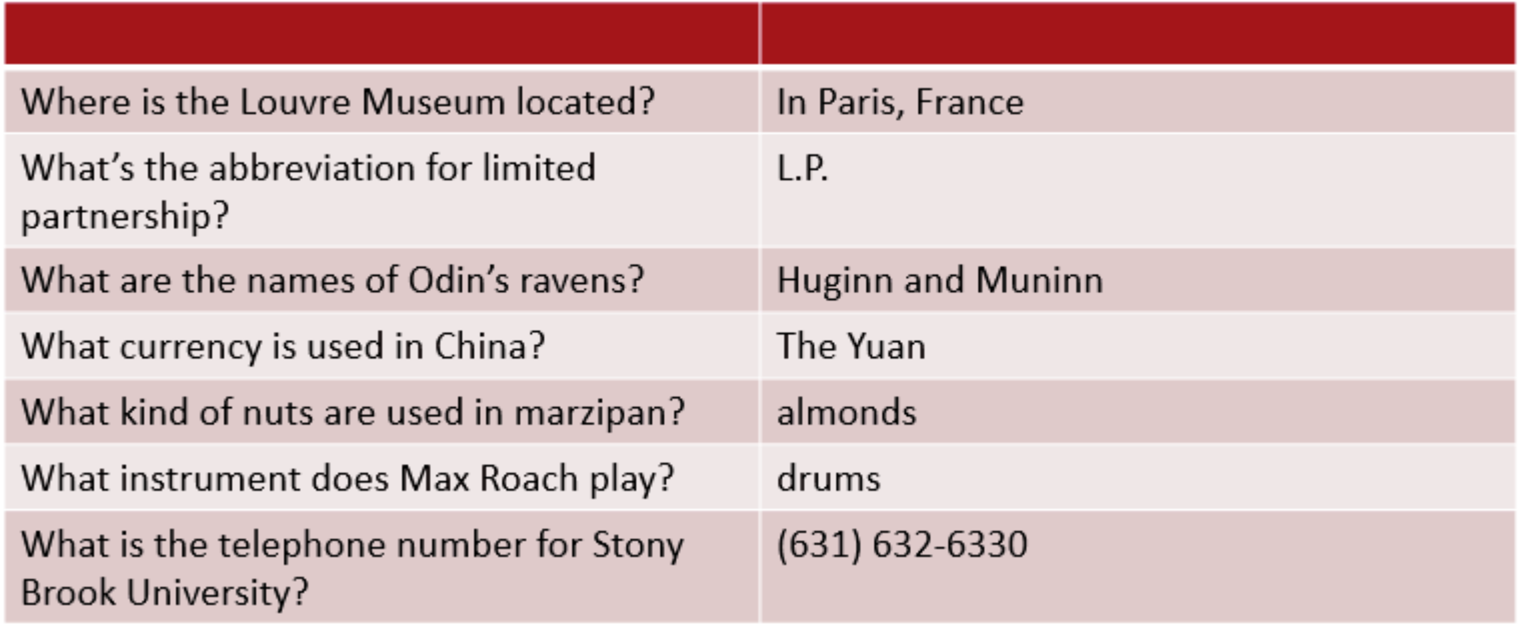
Where is Google based?

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?

Questions for Commercial Systems:



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
What is the telephone number for Stony Brook University?	(631) 632-6330

Paradigms for QA:

IR-based approaches

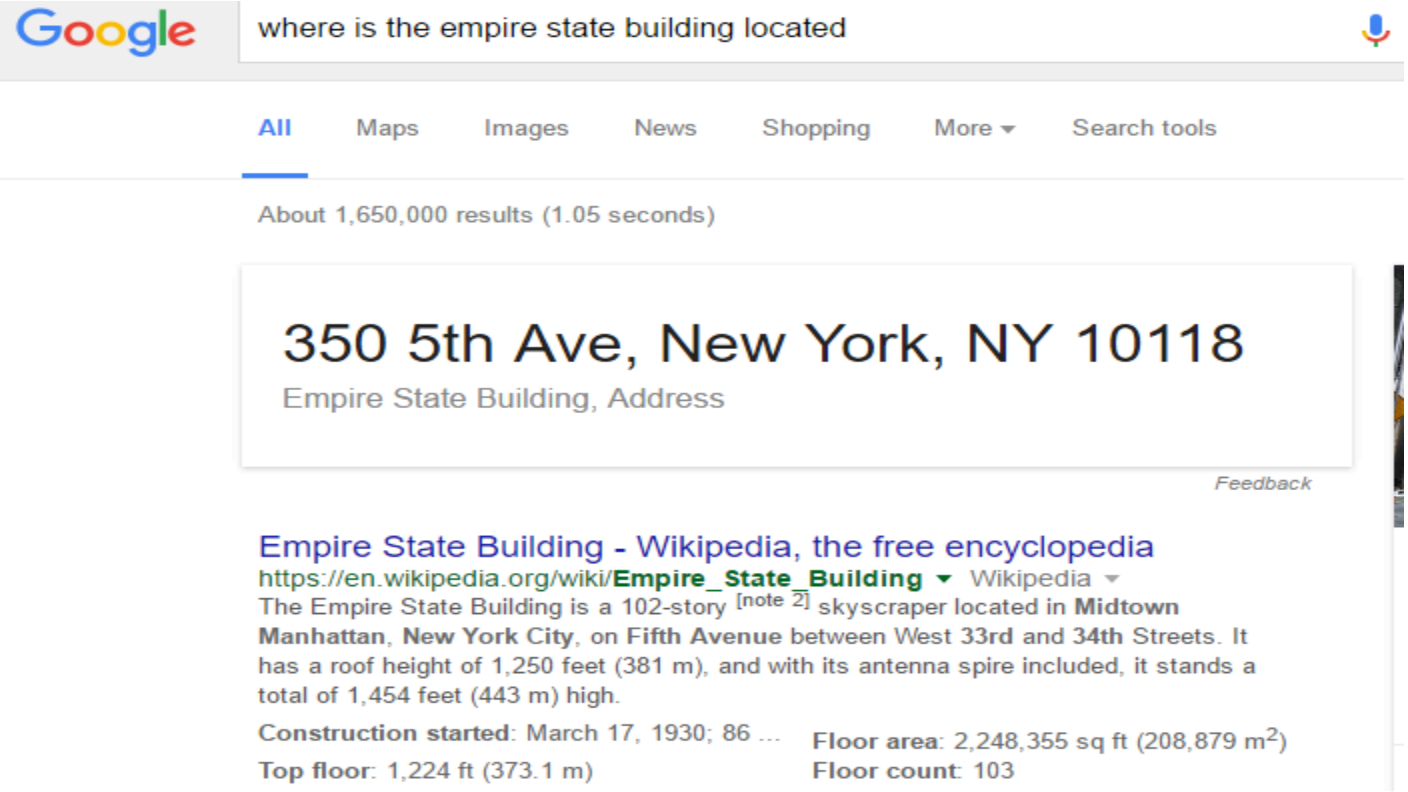
- TREC; IBM Watson; Google

Knowledge-based and Hybrid approaches

- IBM Watson; Apple Siri; Wolfram Alpha; True Knowledge Evi

Reference: Prof. Chris Manning’s Stanford NLP Question Answering Slides

Information Retrieval Based Question Answering:



where is the empire state building located

About 1,650,000 results (1.05 seconds)

350 5th Ave, New York, NY 10118

Empire State Building, Address

Empire State Building - Wikipedia, the free encyclopedia

https://en.wikipedia.org/wiki/**Empire_State_Building** Wikipedia

The Empire State Building is a 102-story [note 2] skyscraper located in **Midtown Manhattan**, **New York City**, on **Fifth Avenue** between West **33rd** and **34th** Streets. It has a roof height of 1,250 feet (381 m), and with its antenna spire included, it stands a total of 1,454 feet (443 m) high.

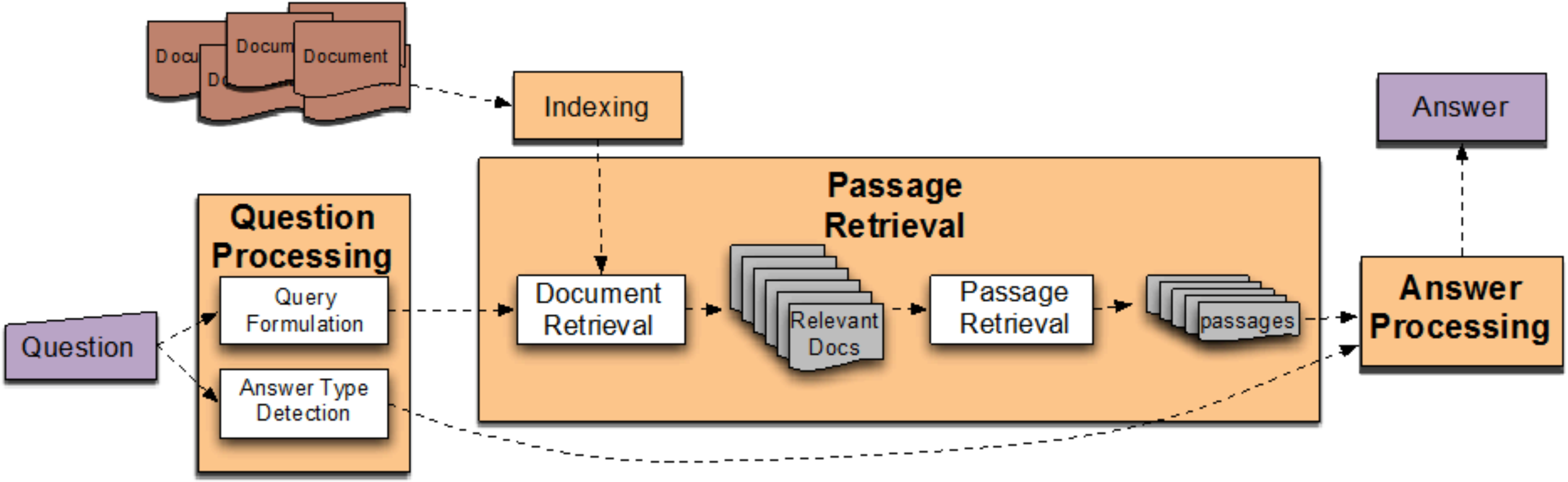
Construction started: March 17, 1930; 86 ...

Top floor: 1,224 ft (373.1 m)

Floor area: 2,248,355 sq ft (208,879 m^2)

Floor count: 103

IR Based QA System:



Reference: Prof. Chris Manning’s Stanford NLP Question Answering Slides

IR Based QA System:

- 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

Answer Type Detection:

Question: What are the two states you could be reentering if you’re crossing New York’s southern border?

Answer Type: US state

Query: two states, border, New York, south

Focus: the two states

Relations: borders(New York, ?x, south)

Answer Type Detection: Named Entities

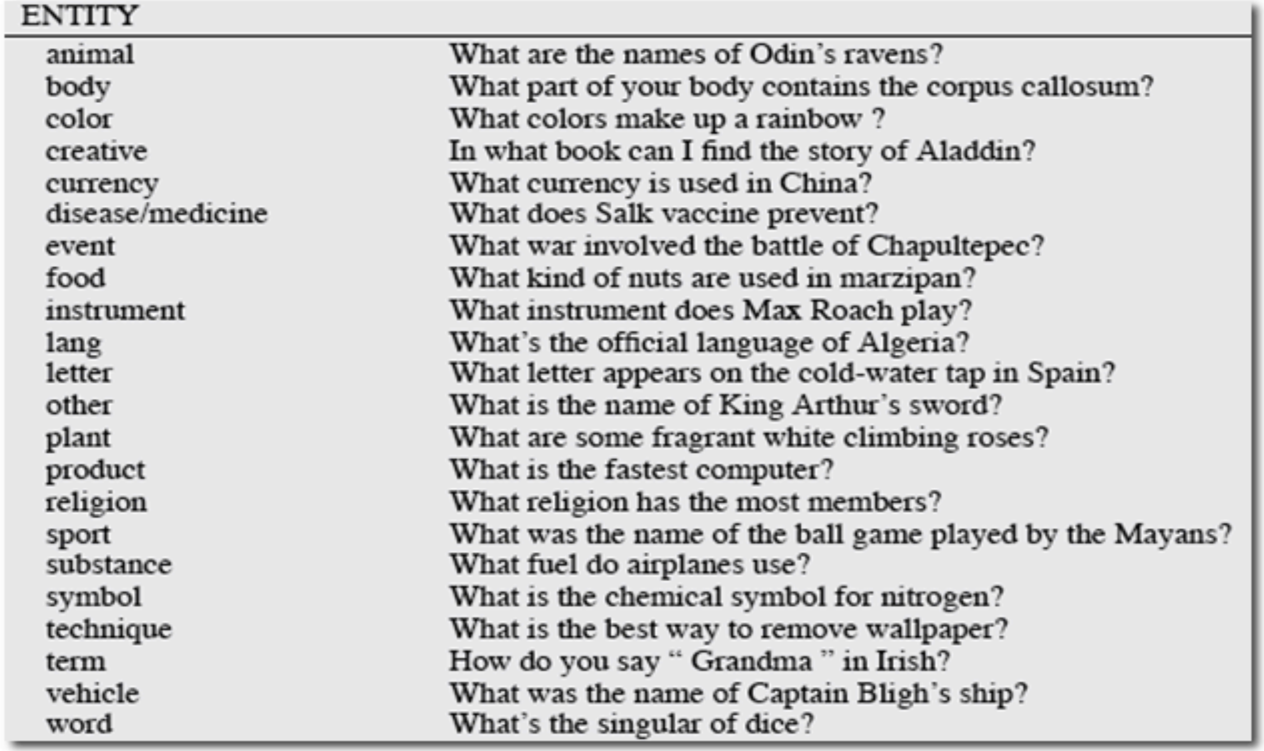
Who founded Microsoft?

- PERSON

Which country in Europe has the largest population?

- *COUNTRY.*

More 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?
food	What kind of nuts are used in marzipan?
instrument	What instrument does Max Roach play?
lang	What's the official language of Algeria?
letter	What letter appears on the cold-water tap in Spain?
other	What is the name of King Arthur's sword?
plant	What are some fragrant white climbing roses?
product	What is the fastest computer?
religion	What religion has the most members?
sport	What was the name of the ball game played by the Mayans?
substance	What fuel do airplanes use?
symbol	What is the chemical symbol for nitrogen?
technique	What is the best way to remove wallpaper?
term	How do you say " Grandma " in Irish?
vehicle	What was the name of Captain Bligh's ship?
word	What's the singular of dice?

Reference: Prof. Chris Manning's Stanford NLP Question Answering Slides

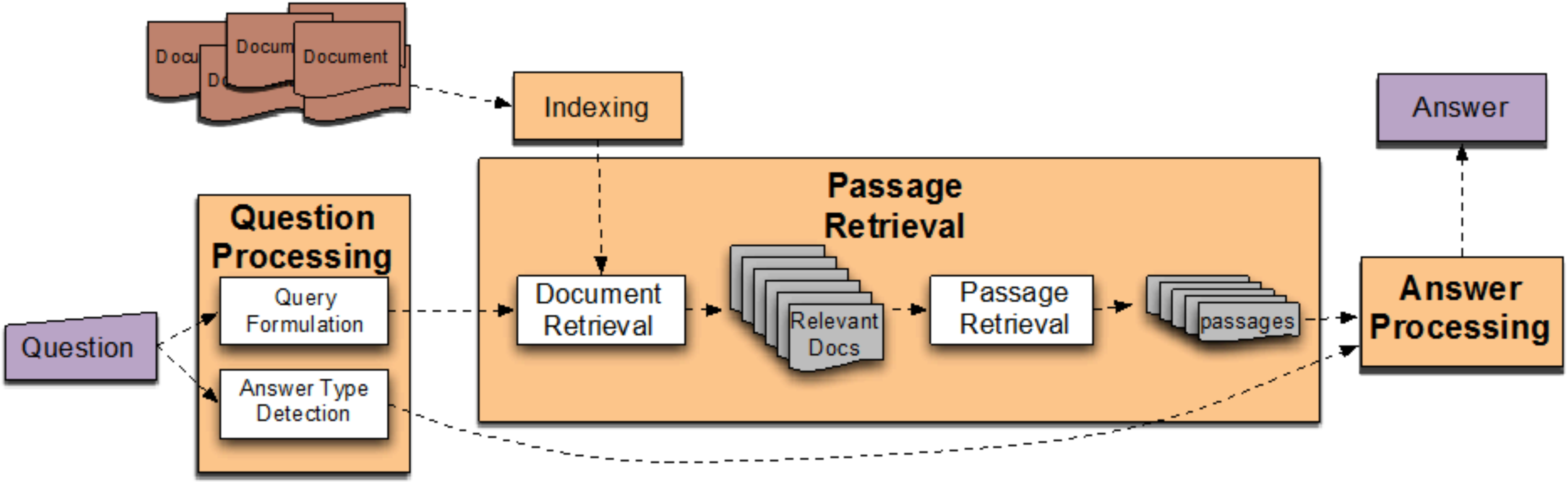
Answer Type Detection:

- Problem treated as a Machine Learning Classification problem.
- 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

Step 2 : Passage Retrieval



Reference: Prof. Chris Manning’s Stanford NLP Question Answering Slides

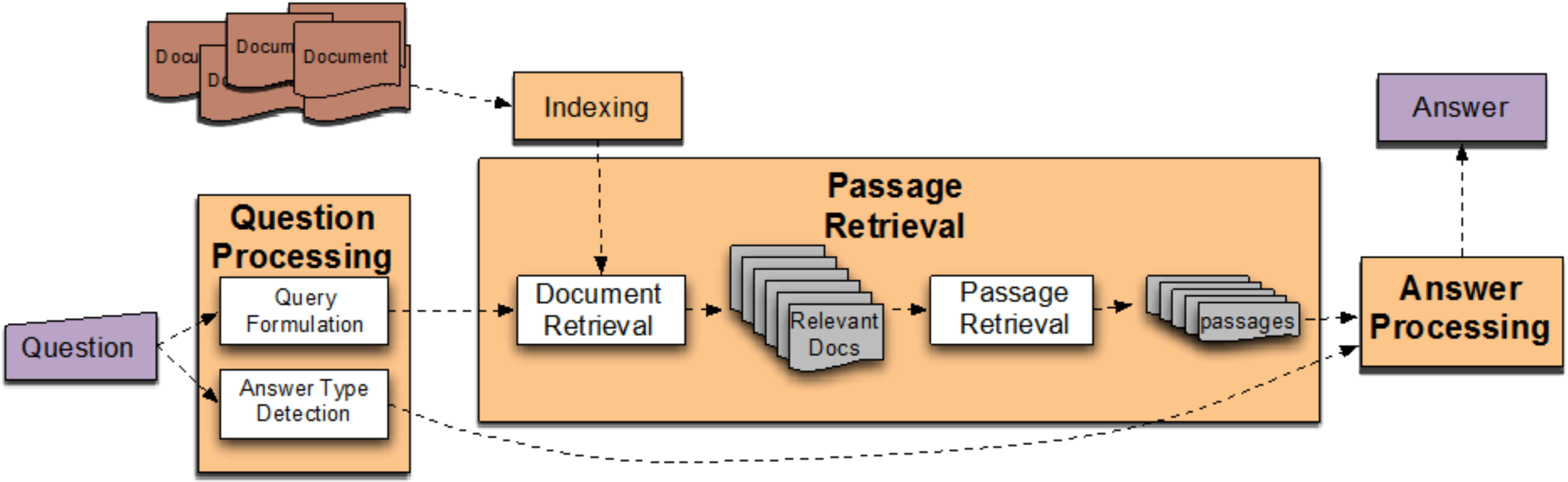
Passage Retrieval:

- 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 re-rank passages.

Features for Passage Ranking:

- 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

Step 3 : Answer Extraction



Reference: Prof. Chris Manning's Stanford NLP Question Answering Slides

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)

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

- How tall is Mt. Everest? (LENGTH)

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

Reference: Prof. Chris Manning's Stanford NLP Question Answering Slides

What if there are multiple Candidate Answers?

Have to rank the 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**

Evaluation Metrics:

1. *Accuracy* (does answer match gold-labeled answer?)
2. *Mean Reciprocal Rank*

- For each query return a ranked list of M candidate answers.
- Query score is 1/Rank of the first correct answer
    - *If first answer is correct: 1*
    - *else if second answer is correct: ½*
    - *else if third answer is correct: ⅓, etc.*
    - *Score is 0 if none of the M answers are correct*
    - Take the mean over all N queries

$$MRR = \frac{\sum_{i=1}^{N} \frac{1}{rank_i}}{N}$$

Reference: Prof. Chris Manning’s Stanford NLP Question Answering Slides

Major Challenges:

- Acquiring high-quality, high-coverage lexical resources
- Improving document retrieval
- Improving document understanding
- Expanding to multi-lingual corpora
- Answer Justification
  - Why should the user trust the answer?
  - Is there a better answer out there?

References:

- Stanford NLP Question Answering Slides http://spark-public.s3.amazonaws.com/nlp/slides/qa.pdf
- CMU NLP Question Answering Slides https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=2&ved=0ahUKEwiY7ZmSkrrMAhWHn4MKHZ_dB8sQFggnMAE&url=http%3A%2F%2Fwww.cs.cmu.edu%2F~ehn%2F15-381%2F15381_QA.ppt&usg=AFQjCNFGeQOKtT2SMIBueaPtrRoHnaYDPw&cad=rja
- https://**opus4**.kobv.de/**opus4**-fau/files/54/chapter03.pdf