Introduction to IR and IE

Vasudeva Varma

IIIT Hyderabad

Approaches to IR

- Two types of retrieval
  - By metadata (subject headings, keywords, etc.)
  - By content
- Metadata as manually assigned information
  - Human agreement is not good
  - Expensive for most data
- Metadata assigned automatically
  - Quality is reasonable, but not high for many applications
- Metadata in general
  - Requires *a priori* prediction of headings, keywords, …
- Most successful IR approaches are content-based

basic approach to IR

- Successful content-based approaches are statistical
  - Rather than actual "understanding" of text
  - Text understanding effectiveness is very poor
    - Exception: works better in some restricted domains
- IR statistics used in different ways
  - Past/concurrent queries and relevance judgments
    - Collaborative systems
  - Document and query similarities

relevant items are similar

- Much of IR depends upon idea that
  similar → relevant to same queries

- Usually measure query-document similarity
  - Can consider document-document similarity

- "Similar" can be measured in many ways
  - String matching
  - Same vocabulary
  - Probability arise from same model
  - Same meaning
  - …

"bag of words"

- An effective and popular approach
- Compares words without regard to order

- Consider reordering words in a headline
  - Stocks fall on inflation fears
  - inflation stocks fall on fears
  - fall inflation stocks on fears
  - fall fears inflation stocks on
  - fall fears inflation on stocks

- Q: How far can we push "bag of words"?

IR engines: State of the Art

- Wide variation in retrieval results
  - User topic
  - Retrieval system
- Different approaches work for different systems.
- No way to determine which approach will work for a particular query.

Solution:

- **Deeper analysis of the content and Query**

Motivation for Deeper Analysis

- Texts are one of the major sources of information and knowledge.

However, they are not transparent.

They have to be systematically integrated with the other sources like data bases, numerical data, etc.

NLP/IR/IE for better analysis
IA for better presentation

IR vs. IE vs. IA

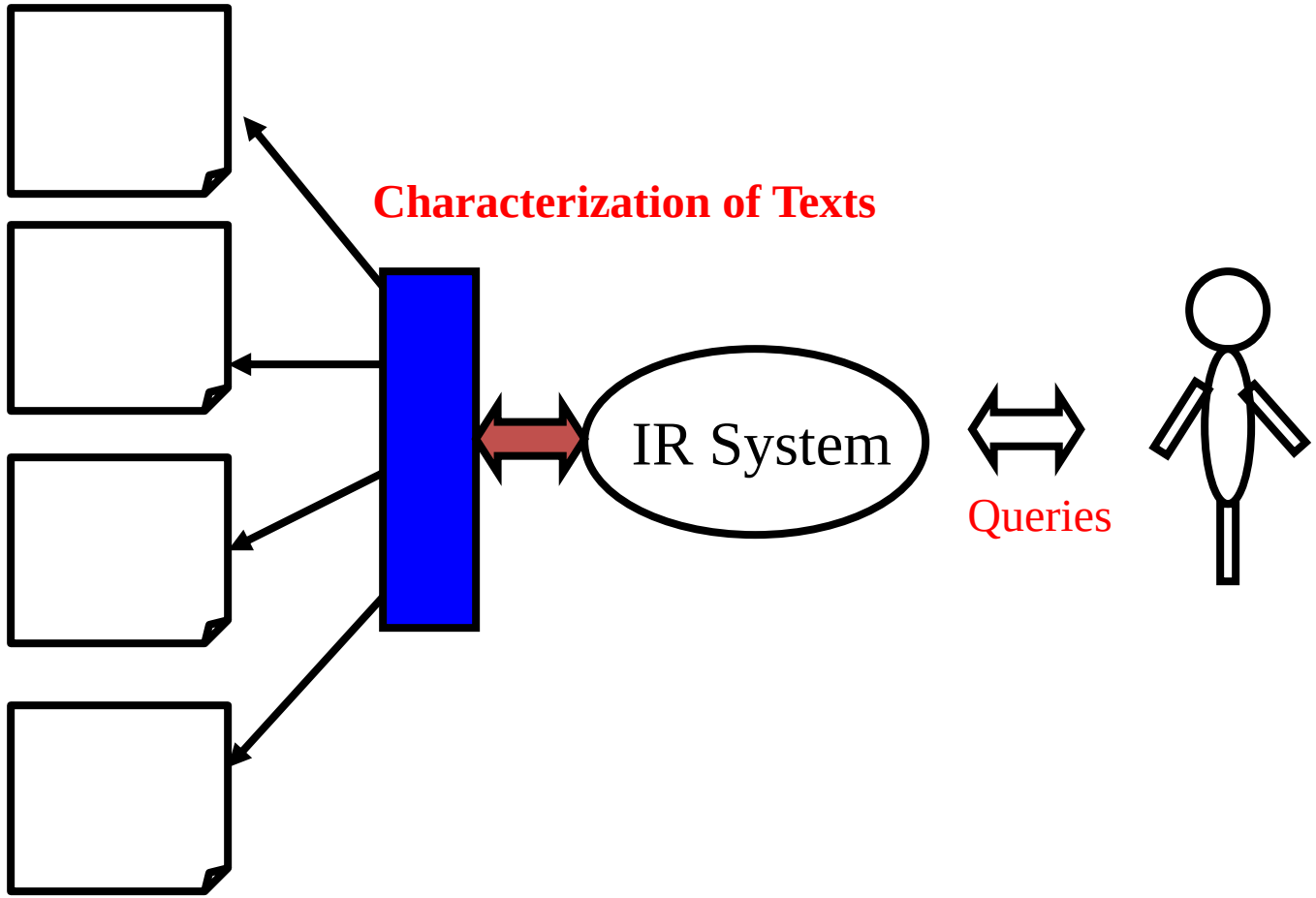
- To search and retrieve documents in response to queries for information

Vs.

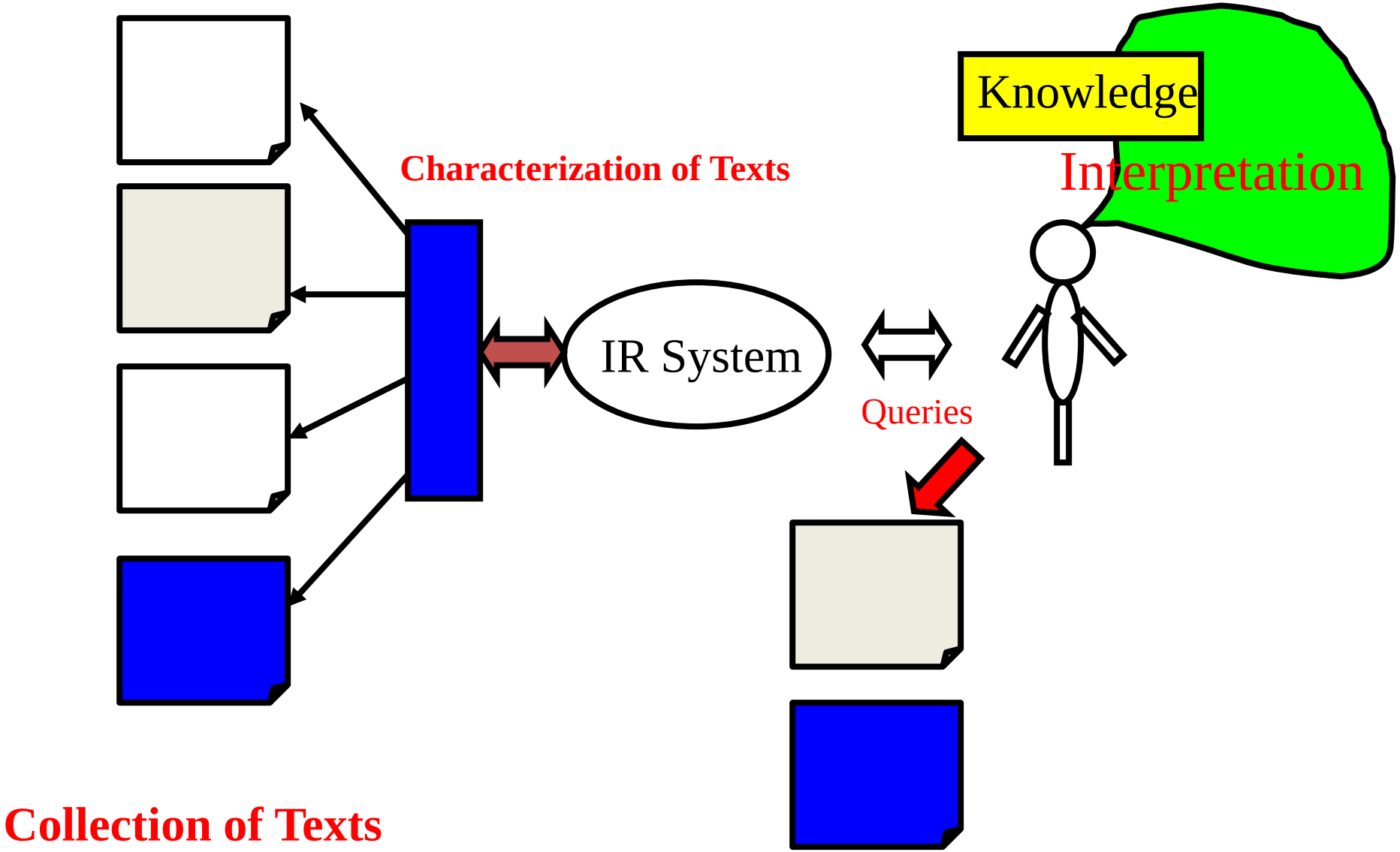
- To extract information that fits pre-defined database schemas or templates, specifying the output formats

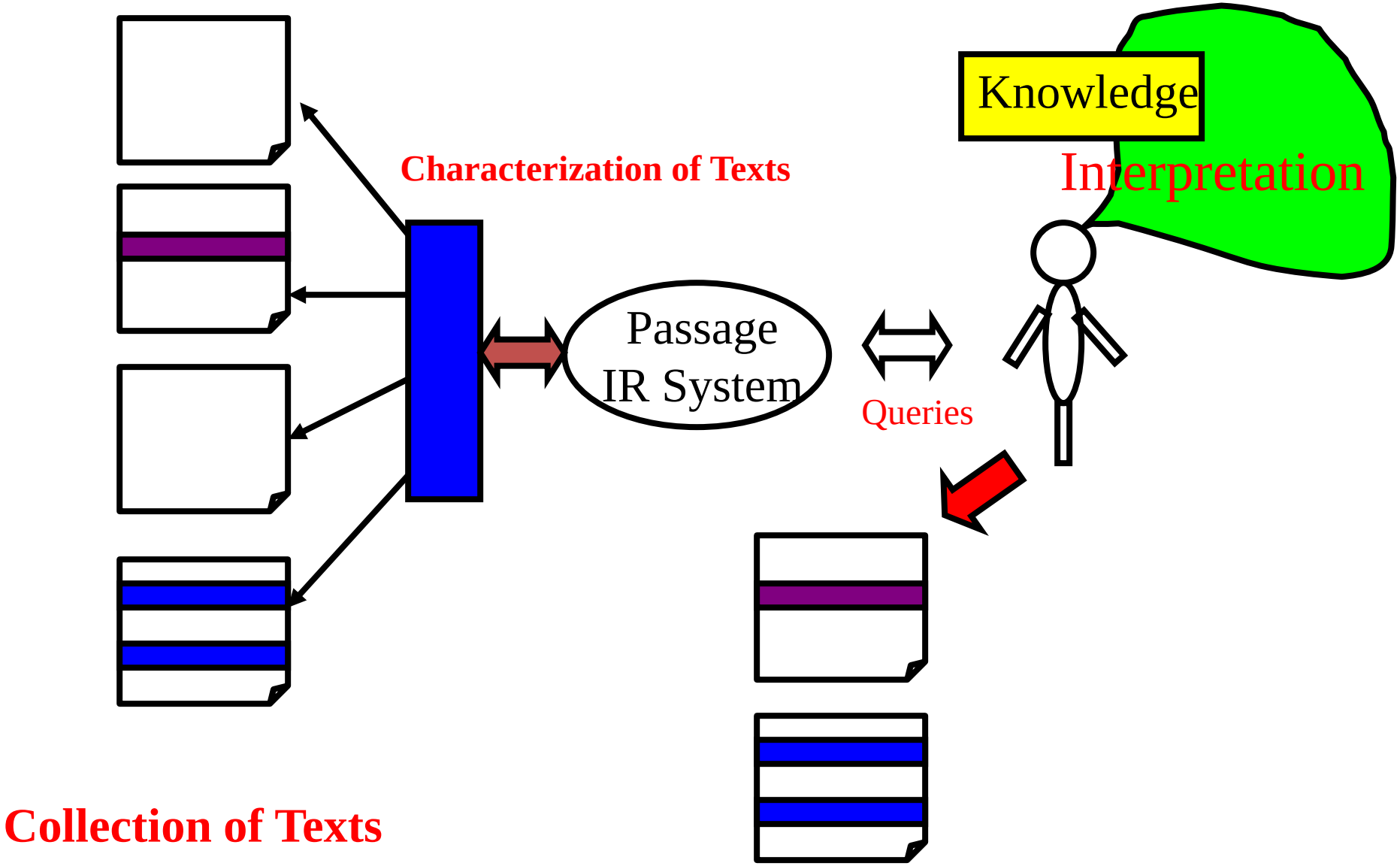
Vs.

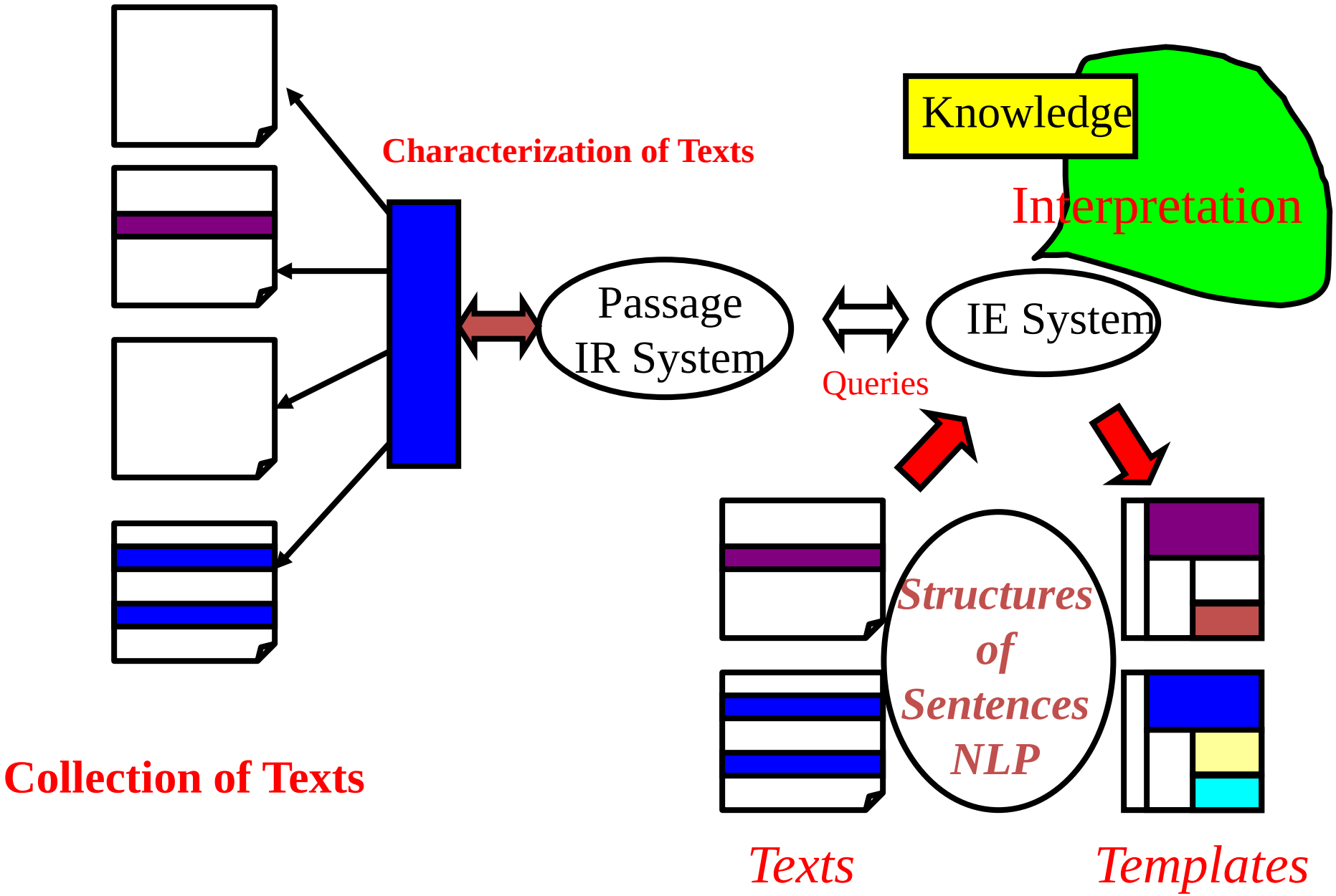
- To make the required information accessible to the user in their choice of language, mode, level of detail and format

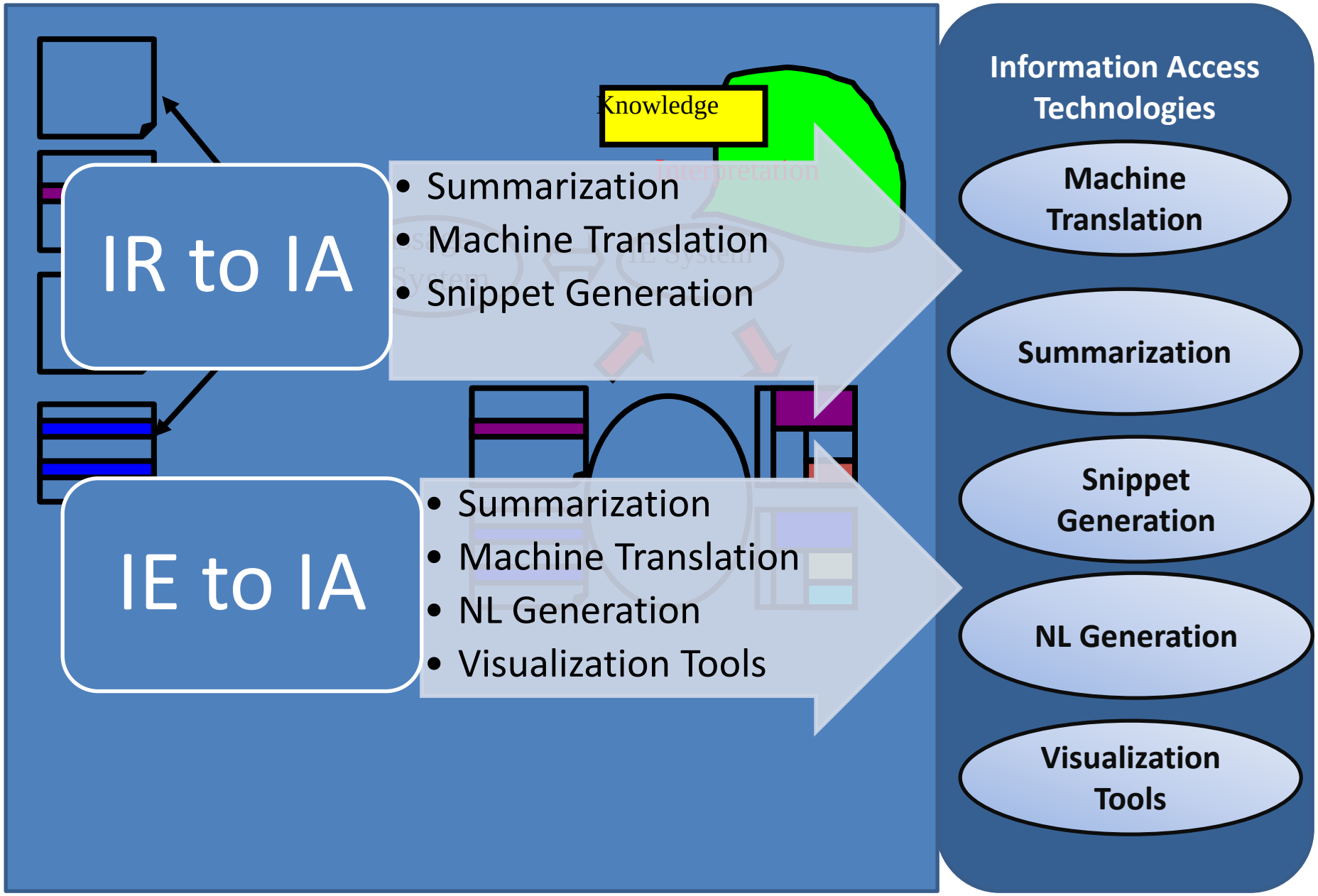


Collection of Texts


Knowledge
Interpretation
Characterization of Texts
IR System
Queries
Collection of Texts


Knowledge
Interpretation
Characterization of Texts
Passage
IR System
Queries
Collection of Texts


Characterization of Texts
Knowledge
Interpretation
Passage
IR System
IE System
Queries
Structures
of
Sentences
NLP
Collection of Texts
Texts
Templates


Knowledge
IR to IA
• Summarization
• Machine Translation
• Snippet Generation
IE to IA
• Summarization
• Machine Translation
• NL Generation
• Visualization Tools
Information Access Technologies
Machine Translation
Summarization
Snippet Generation
NL Generation
Visualization Tools

how do we represent text?

- Remember: computers don't "understand" anything!
- "Bag of words"
  - Treat all the words in a document as index terms
  - Assign a "weight" to each term based on "importance" (or, in simplest case, presence/absence of word)
  - Disregard order, structure, meaning, etc. of the words
  - Simple, yet effective!
- Assumptions
  - Term occurrence is independent
  - Document relevance is independent
  - "Words" are well-defined

what's a word?

天主教教宗若望保祿二世因感冒再度住進醫院。
這是他今年第二度因同樣的病因住院。

الناطق باسم -وقال مارك ريجيف
إن شارون قبل -الخارجية الإسرائيلية
الدعوة وسيقوم للمرة الأولى بزيارة
تونس، التي كانت لفترة طويلة المقر
.1982الرسمي لمنظمة التحرير الفلسطينية بعد خروجها من لبنان عام

Выступая в Мещанском суде Москвы экс-глава ЮКОСа заявил не совершал ничего противозаконного, в чем обвиняет его генпрокуратура России.

भारत सरकार ने आर्थिक सर्वेक्षण में वित्तीय वर्ष 2005-06 में सात फ़ीसदी विकास दर हासिल करने का आकलन किया है और कर सुधार पर ज़ोर दिया है

日米連合で台頭中国に対処…アーミテージ前副長官提言

조재영 기자= 서울시는 25일 이명박 시장이 `행정중심복합도시" 건설안에 대해 `군대라도 동원해 막고싶은 심정"이라고 말했다는 일부 언론의 보도를 부인했다.

sample Document

McDonald's slims down spuds

Fast-food chain to reduce certain types of fat in its french fries with new cooking oil.

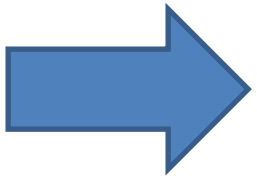
NEW YORK (CNN/Money) - McDonald's Corp. is cutting the amount of "bad" fat in its french fries nearly in half, the fast-food chain said Tuesday as it moves to make all its fried menu items healthier.

But does that mean the popular shoestring fries won't taste the same? The company says no. "It's a win-win for our customers because they are getting the same great french-fry taste along with an even healthier nutrition profile," said Mike Roberts, president of McDonald's USA.

But others are not so sure. McDonald's will not specifically discuss the kind of oil it plans to use, but at least one nutrition expert says playing with the formula could mean a different taste.

Shares of Oak Brook, Ill.-based McDonald's (MCD: down $0.54 to $23.22, Research, Estimates) were lower Tuesday afternoon. It was unclear Tuesday whether competitors Burger King and Wendy's International (WEN: down $0.80 to $34.91, Research, Estimates) would follow suit. Neither company could immediately be reached for comment.

…



“Bag of Words”

14 × McDonalds

12 × fat

11 × fries

8 × new

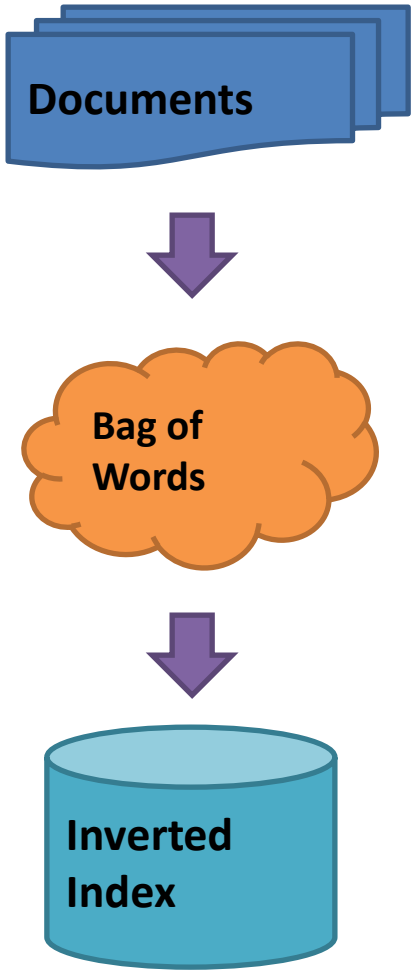
7 × french

6 × company, said, nutrition

5 × food, oil, percent, reduce, taste, Tuesday

…

counting words…



case folding, tokenization, stop word removal, stemming

Retrieval Models

- Boolean
- Vector space
  - Basic vector space — *SMART, LUCENE*
  - Extended Boolean
- Probabilistic models
  - Statistical language models — *Lemur*
  - Two Poisson model — *Okapi*
  - Bayesian inference networks — *Inquery*
- Citation/Link analysis models
  - Page rank — *Google*
  - Hub & authorities — *Clever*

Types of Retrieval Models

- Exact Match (Document Selection)
  - Example: Boolean Retrieval Method
  - Query defines the exact retrieval criterion
  - Relevance is a binary variable; a document is either relevant (i.e., match query) or irrelevant (i.e., mismatch)
  - Result is a set of documents
    - Documents are unordered
    - Often in reverse-chronological order (e.g., Pubmed)

Types of Retrieval Models

- Best Match (Document Ranking)
  - Example: Most probabilistic models
  - Query describes the desired retrieval criterion
  - Degree of relevance is a continuous/integral variable; each document matches query to some degree
  - Result in a ranked list (top ones match better)
  - Often return a partial list (e.g., rank threshold)

boolean model

- Simple model based on set theory
- Queries specified as Boolean expressions
  - precise semantics
  - neat formalism
  - $q = ka \wedge (kb \vee \neg kc)$
- Terms are either present or absent.
  - Thus, $w_{ij} \varepsilon \{0,1\}$

term-document incidence matrix

	Anthony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth	...
ANTHONY	1	1	0	0	0	1	
BRUTUS	1	1	0	1	0	0	
CAESAR	1	1	0	1	1	1	
CALPURNIA	0	1	0	0	0	0	
CLEOPATRA	1	0	0	0	0	0	
MERCY	1	0	1	1	1	1	
WORSER	1	0	1	1	1	0	

- Entry is 1 if term occurs. Example: Calpurnia occurs in Julius Caesar.
- Entry is 0 if term doesn't occur. Example: Calpurnia doesn't occur in The tempest.
- We will return to this matrix many times