Language and Computation

week 13, Tuesday, April 22

Tamás Biró
Yale University
tamas.biro@yale.edu
http://www.birot.hu/courses/2014-LC/

Practical matters

- **"Superficial" reading:** JM 22-24
- **Pre-reading:** Intro to JM 25
- **Assignments** 4 returned, 5 posted
- **Python:** if needed, programming section
- **Sections**

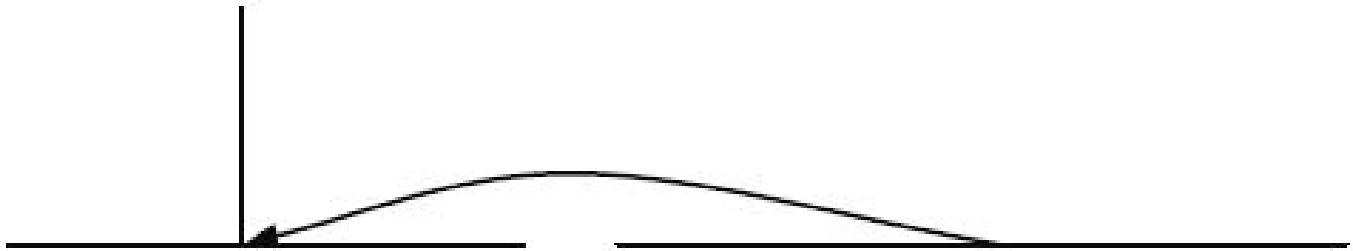
Today

- Anaphora resolution and coreference resolution
  $\rightarrow$ machine learning from vectors
- Information extraction
- Named entity recognition
  $\rightarrow$ machine learning from vectors
- Question answering
  $\rightarrow$ machine learning from vectors

Next time: Machine translation as a summary

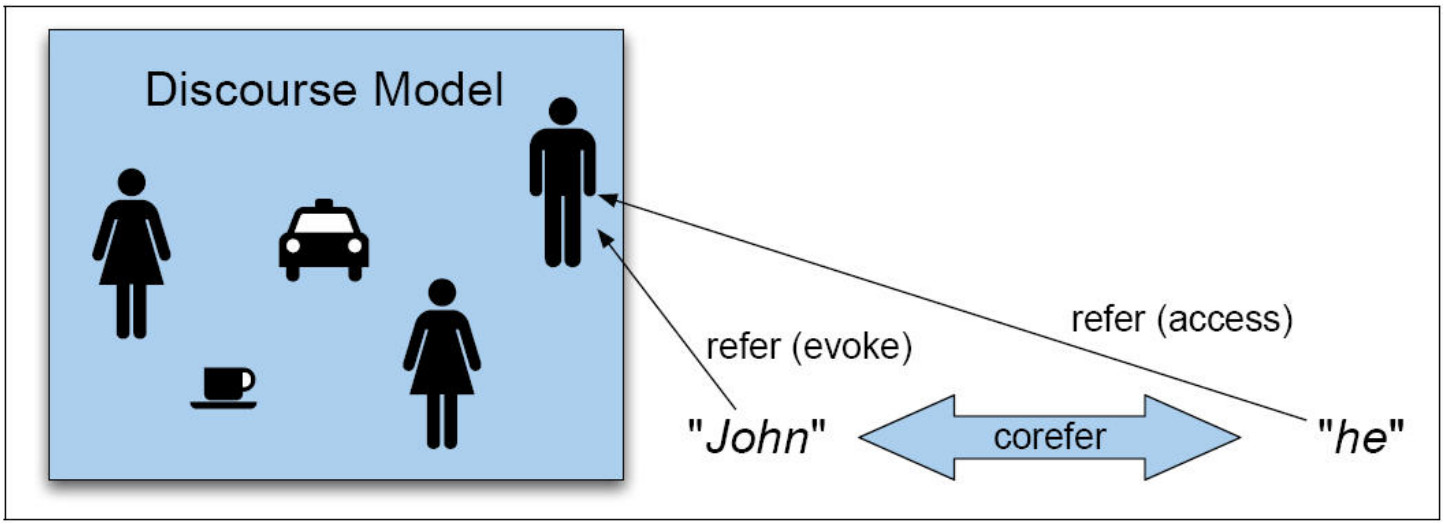
Anaphora resolution, coreference resolution

Anaphora resolution, coreference resolution



$John_i$ *is a good friend of* $Kevin_j$. $He_{i/j?}$ *loves Mary.*

Anaphora resolution, coreference resolution



Anaphora resolution, Coreference resolution

- John Smith$_i$ is a professor. Mr. Smith$_i$ works at Yale.
- My neighbors are John$_i$ and Marry$_j$.
  He$_i$ is a doctor and she$_j$ is a lawyer.
- Last night I walked my dog$_i$. Max$_i$ was very happy.
- I saw a cat$_i$. The feline$_i$ was black.
- He had a hammer$_i$ with him. The tool$_i$ was heavy.
- Row, row, row your boat, // Gently down the stream.
  Merrily, merrily, merrily, merrily, // Life is but a dream.

Anaphora resolution, Coreference resolution

- Explicit representation of the world, of the context combined with a deep semantic analysis.
- Heuristic approaches (such as the *Hobbs Algorithm*):
  - Take the closest plausible NP in the context.
  - "Closest": take some syntactic information into account (parse tree)
  - Constraints: gender, person, number, binding theory, etc.
- Machine learning

Anaphora resolution, Coreference resolution

Machine learning: (same idea as your midterm)

- $X = \{$ possible anaphora-antecedent pairs $\}$.
- Binary (boolean) classification: $Y = \{t, f\}$.
- Supervised, or unsupervised, or semi-supervised.
- Anaphora-antecedent pairs represented as vectors.
- Classifiers such as log-linear models, Naive Bayes, etc.
- **Evaluation**: precision, recall, f-score.

Features for Pronominal Anaphora Resolution

John saw a beautiful 1961 Ford Falcon at the used car dealership.(U_1)

He showed it to Bob.(U_2)

He bought it.(U_3)

	He (U_2)	it (U_2)	Bob (U_2)	John (U_1)
strict number	1	1	1	1
compatible number	1	1	1	1
strict gender	1	0	1	1
compatible gender	1	0	1	1
sentence distance	1	1	1	2
Hobbs distance	2	1	0	3
grammatical role	subject	object	PP	subject
linguistic form	pronoun	pronoun	proper	proper

Information Extraction

Information Extraction

- Named Entity Recognition
- Event detection
- Relation detection and classification:
  - semantic relations among named entities
  - temporal analysis of events

Extracted information $\rightarrow$ template filling.

Templates representing world/situation/context

Citing high fuel prices, United Airlines said Friday it has increased fares by $6 per round trip on flights to some cities also served by lower-cost carriers. American Airlines, a unit of AMR Corp., immediately matched the move, spokesman Tim Wagner said. United said the increase too effect Thursday [...]

FARE-RAISE ATTEMPT:	LEAD AIRLINE:	UNITED AIRLINES
	AMOUNT:	$6
	EFFECTIVE DATE:	2006-10-26
	FOLLOWER:	AMERICAN AIRLINES

Named Entity Recognition

Generic named entity types:

Type	Tag	Sample Categories
People	PER	Individuals, fictional characters, small groups
Organization	ORG	Companies, agencies, political parties, religious groups, sports teams
Location	LOC	Physical extents, mountains, lakes, seas
Geo-Political Entity	GPE	Countries, states, provinces, counties
Facility	FAC	Bridges, buildings, airports
Vehicles	VEH	Planes, trains, and automobiles

Type	Example
People	*Turing* is often considered to be the father of modern computer science.
Organization	The *IPCC* said it is likely that future tropical cyclones will become more intense.
Location	The *Mt. Sanitas* loop hike begins at the base of *Sunshine Canyon*.
Geo-Political Entity	*Palo Alto* is looking at raising the fees for parking in the University Avenue district.
Facility	Drivers were advised to consider either the *Tappan Zee Bridge* or the *Lincoln Tunnel*.
Vehicles	The updated *Mini Cooper* retains its charm and agility.

Named Entity Recognition

Ambiguities everywhere:

Name	Possible Categories
Washington	Person, Location, Political Entity, Organization, Facility
Downing St.	Location, Organization
IRA	Person, Organization, Monetary Instrument
Louis Vuitton	Person, Organization, Commercial Product

[$_{PERS}$ 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.
The [$_{FAC}$ Washington] had proved to be a leaky ship, every passage I made...

Named Entity Recognition

Potential named entities as feature vectors:

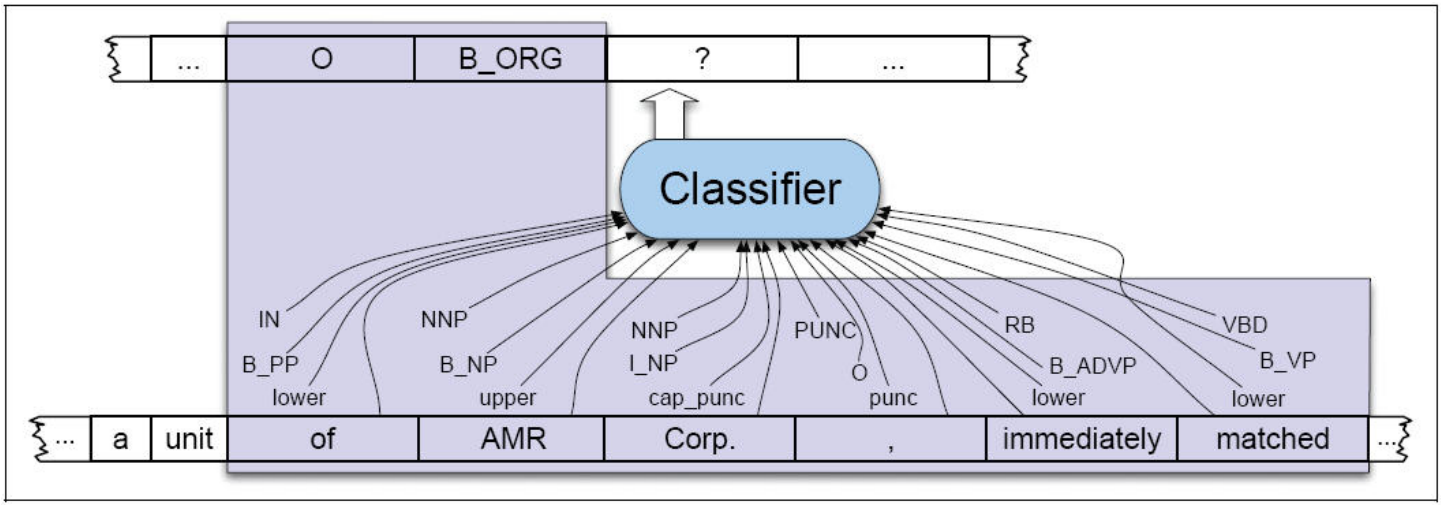
Feature	Explanation
Lexical items	The token to be labeled
Stemmed lexical items	Stemmed version of the target token
Shape	The orthographic pattern of the target word
Character affixes	Character-level affixes of the target and surrounding words
Part of speech	Part of speech of the word
Syntactic chunk labels	Base-phrase chunk label
Gazetteer or name list	Presence of the word in one or more named entity lists
Predictive token(s)	Presence of predictive words in surrounding text
Bag of words/Bag of N-grams	Words and/or *N*-grams occurring in the surrounding context

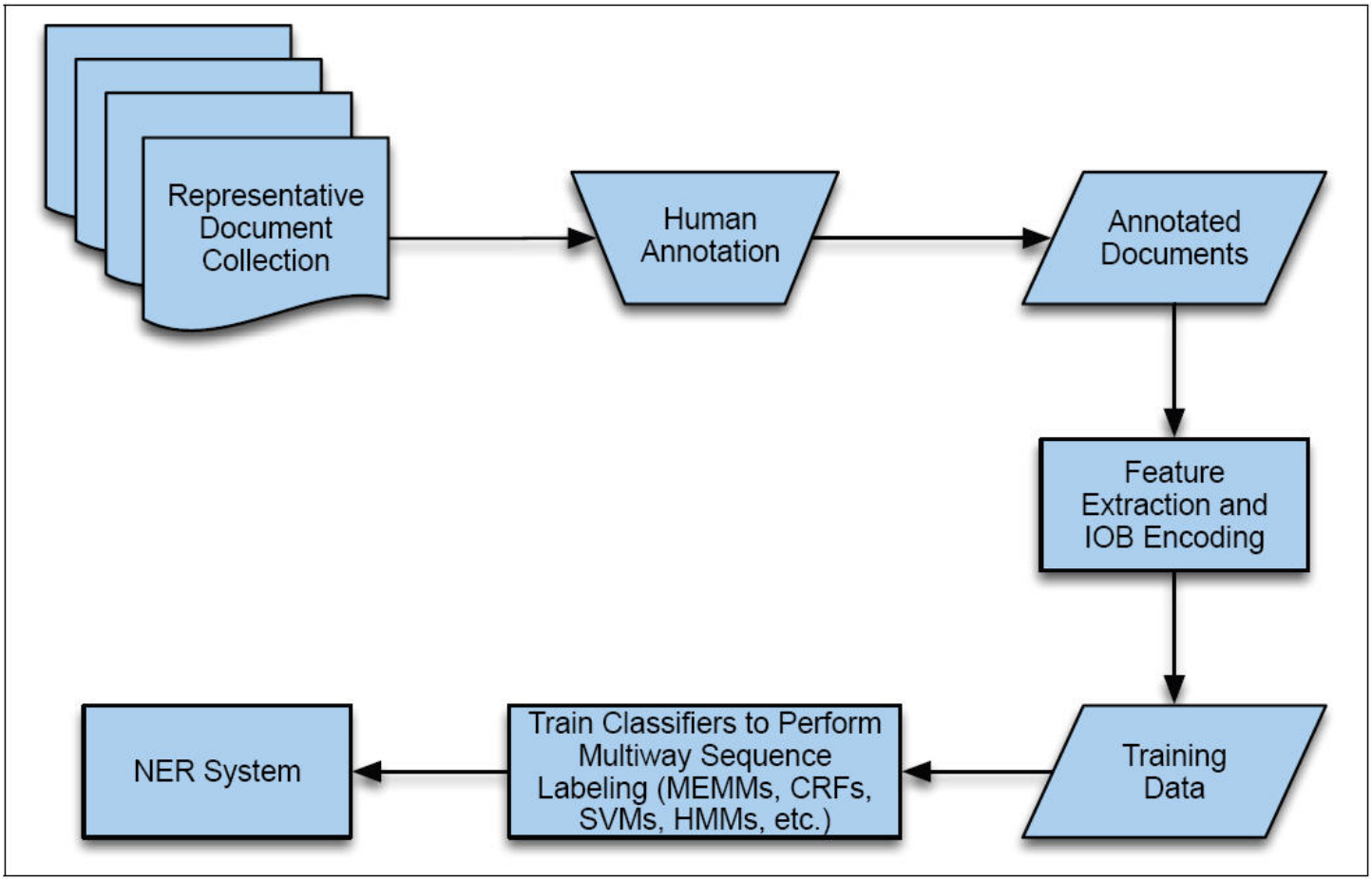
Shape	Example
Lower	cummings
Capitalized	Washington
All caps	IRA
Mixed case	eBay
Capitalized character with period	H.
Ends in digit	A9
Contains hyphen	H-P

Named Entity Recognition

Machine learning: (same idea as your midterm)

- $X = \{$ possible named entities $\}$.
- Binary classification: $Y = \{t, f\}$, for each NER type.
- Supervised, or unsupervised, or semi-supervised.
- Possible named entities represented as vectors.
- Classifiers such as log-linear models, Naive Bayes, etc.
- **Evaluation**: precision, recall, f-score.


...
O
B_ORG
?
...
Classifier
IN
B_PP
lower
NNP
B_NP
upper
NNP
I_NP
cap_punc
PUNC
O
punc
RB
B_ADVP
lower
VBD
B_VP
lower
...
a
unit
of
AMR
Corp.
,
immediately
matched
...

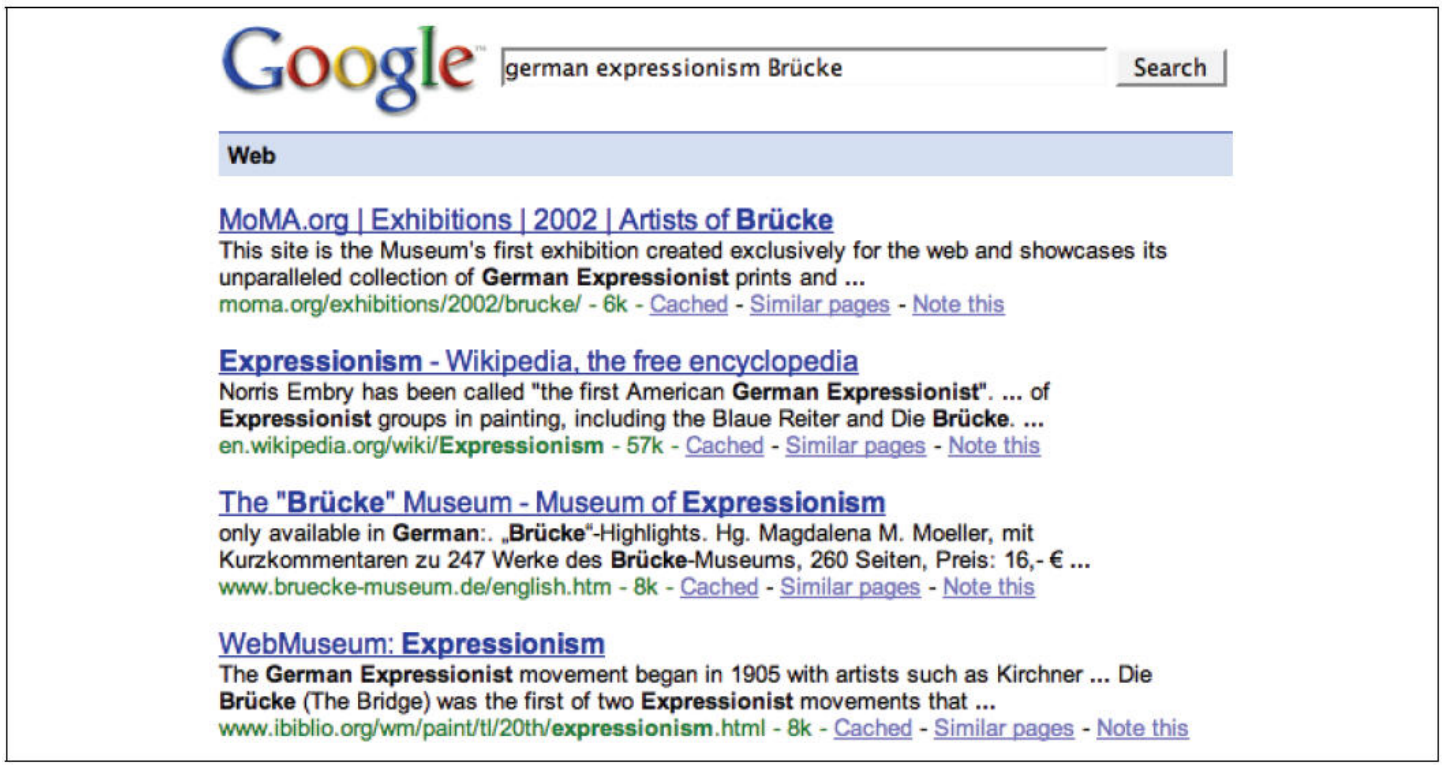

Representative Document Collection
Human Annotation
Annotated Documents
Feature Extraction and IOB Encoding
Training Data
Train Classifiers to Perform Multiway Sequence Labeling (MEMMs, CRFs, SVMs, HMMs, etc.)
NER System

Question Answering and Summarization

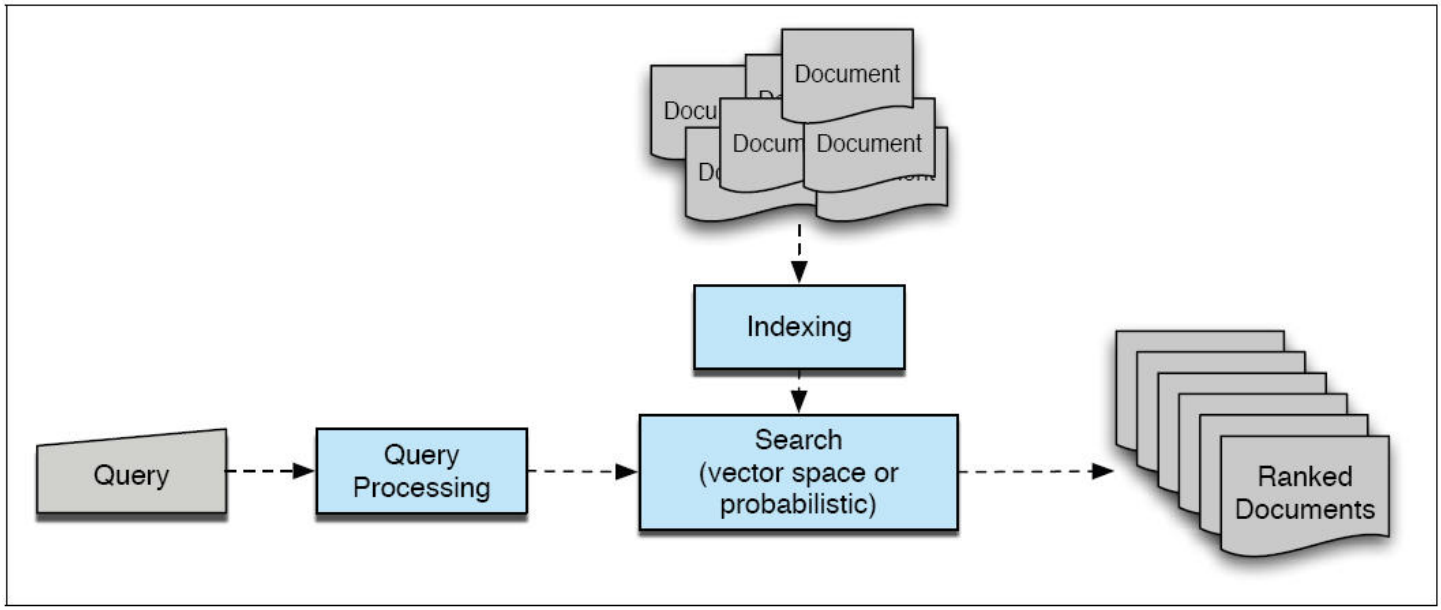
Question Answering and Summarization

- **Information Retrieval (IR):** return documents that are relevant to a particular natural language query.
- (Passage retrieval)
- **Question Answering (QA):** find answer in the documents (a word, a phrase, a sentence)
- **Text summarization:** produce an abridged version
  Includes *natural language generation*.

IR: Information Retrieval



IR: Information Retrieval



Question Answering: passage retrieval

Google when was movable type metal printing invented in ko Search

Web Results 1 -

Movable type - Wikipedia, the free encyclopedia
Metal movable type was first **invented in Korea** during the Goryeo Dynasty .... oldest extant **movable metal print** book is the Jikji, printed in **Korea** in 1377. ...
en.wikipedia.org/wiki/**Movable_type** - 78k - Cached - Similar pages - Note this

Hua Sui - Wikipedia, the free encyclopedia
Hua Sui is best known for creating China's first **metal movable type printing** in 1490 AD. **Metal movable type printing** was also **invented in Korea** during the ...
en.wikipedia.org/wiki/Hua_Sui - 40k - Cached - Similar pages - Note this
[More results from en.wikipedia.org]

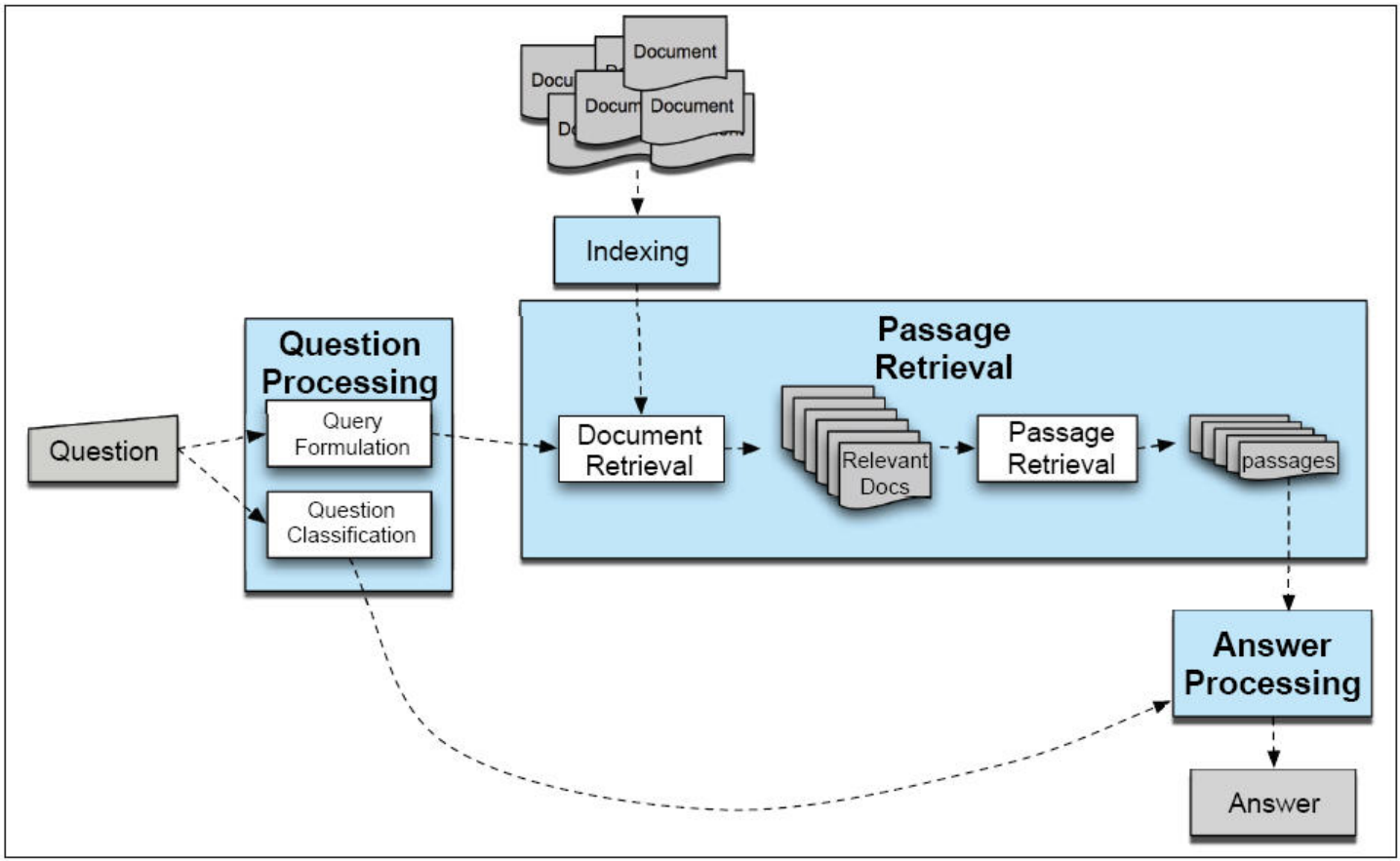
Education and Literacy
Korea has a long and venerable tradition of **printing** and publishing. In particular it can boast the world's first serious use of **movable metal type** in ...
mmtaylor.net/Literacy_Book/DOCS/16.html - 8k - Cached - Similar pages - Note this

Earliest Printed Books in Select Languages, Part 1: 800-1500 A.D. ...
This is the oldest extant example of **movable metal type printing**. **Metal type** was used in **Korea** as early as 1234; in 1403 King Htai Tjong ordered the first ...
blogs.britannica.com/blog/main/2007/03/earliest-printed-books-in-selected-languages-part-1-800-1500-ad/ - 47k - Cached - Similar pages - Note this

QA: Factoid Question Answering

Question	Answer
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's the official language of Algeria?	Arabic
What is the telephone number for the University of Colorado, Boulder?	(303)492-1411
How many pounds are there in a stone?	14

QA: Factoid Question Answering


Document
Document
Indexing
Question Processing
Query Formulation
Question Classification
Question
Passage Retrieval
Document Retrieval
Relevant Docs
Passage Retrieval
passages
Answer Processing
Answer

Entailment

- **Upward entailment:**
  She sang in French. $\Rightarrow$ *She sang.*

- **Downward entailment:**
  She did not sing in French. $\Leftarrow$ *She did not sing.*

- **No entailment:**
  Exactly three students sang in French.
  vs. *Exactly three students sang.*

General recipe

Machine learning: (same idea as your midterm)

- $X = \{$ possible . . . $\}$.
- Binary classification: $Y = \{t, f\}$.
- Supervised, or unsupervised, or semi-supervised.
- . . . represented as vectors.
- Classifiers such as log-linear models, Naive Bayes, etc.
- **Evaluation**: precision, recall, f-score.

See you next week!