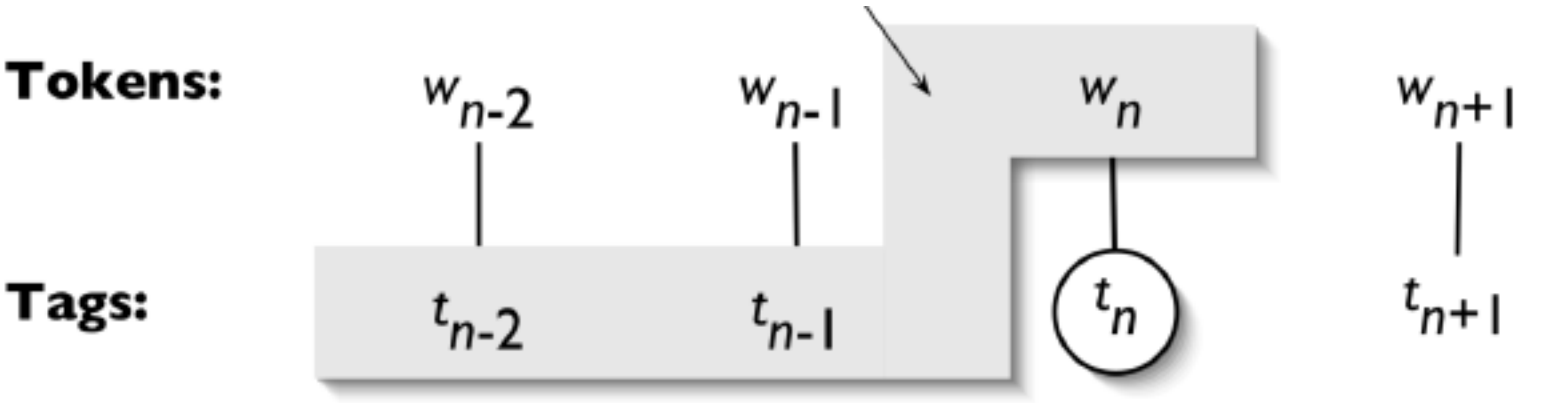
COMP90042
Web search and text analysis

Workshop Week 8

xudong.han@unimelb.edu.au
https://github.com/HanXudong/COMP90042_Workshops

What are some common approaches to POS tagging? What aspects of the data might allow us to predict POS tags systematically?

- N-gram
- Rule-based https://learnenglish.britishcouncil.org/english-grammar
- Classifier



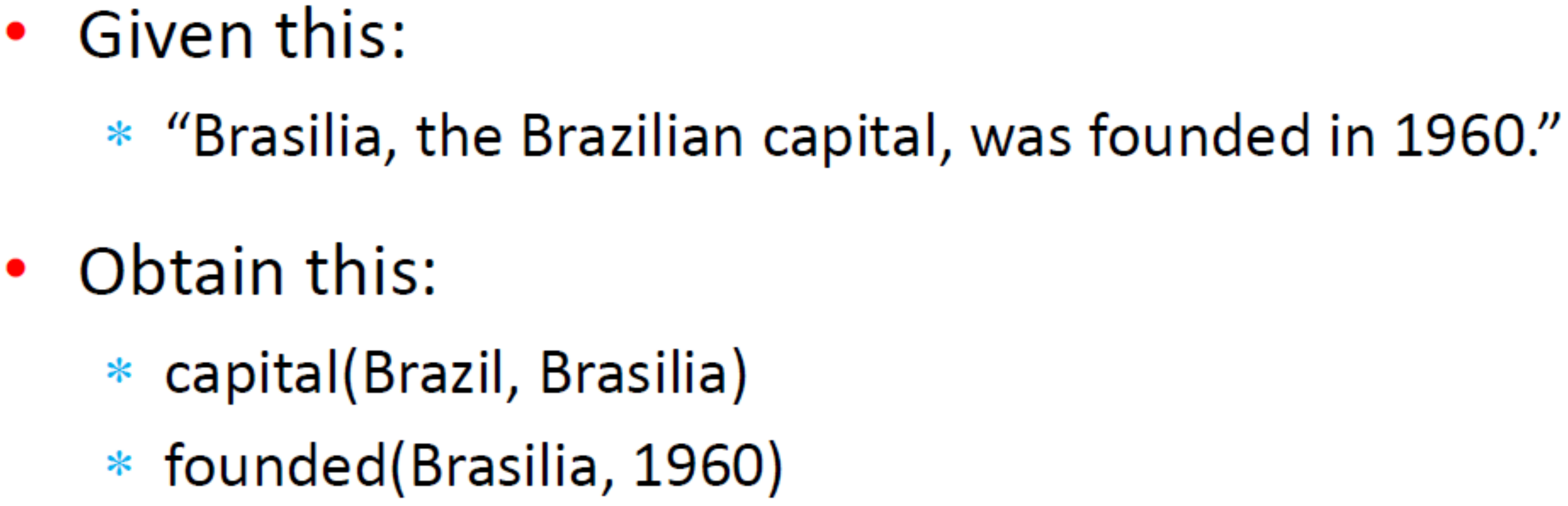
- HMM

At the end of this tutorial you will be able to…

1. explain the basic idea of information extraction
    1. explain what is Named Entity and how to do NER
2. tell the differences between IO and IOB methods
3. explain what is semantic parsing
4. build a simple QA

Q1
What is Information Extraction?
What might the “extracted” information look like?

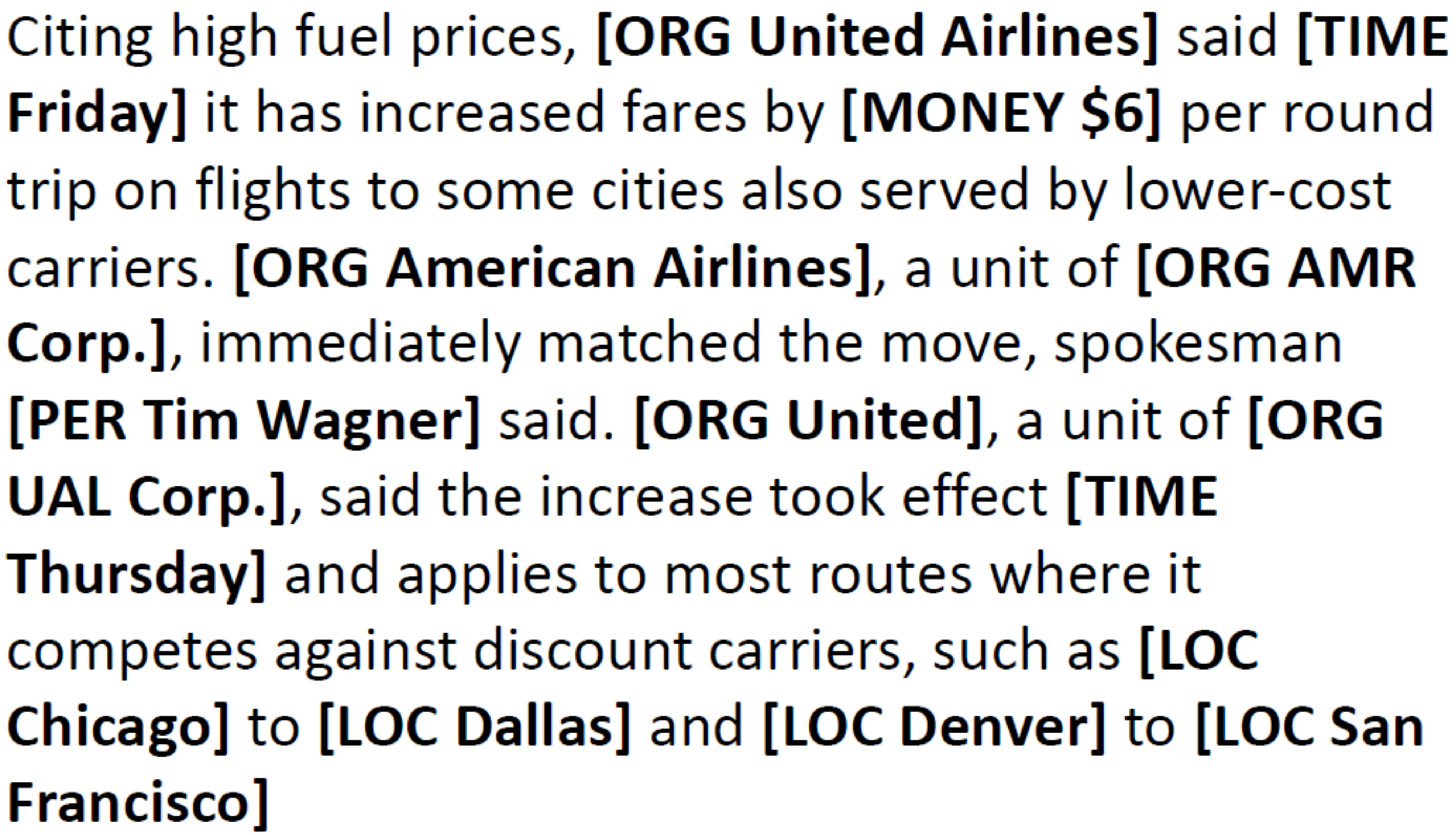
- extract information into a structured format
- Structured: databases …
- Unstructured: text



- Given this:
  - * “Brasilia, the Brazilian capital, was founded in 1960.”
- Obtain this:
  - * capital(Brazil, Brasilia)
  - * founded(Brasilia, 1960)

What is Named Entity Recognition and why is it difficult?

- In information extraction, a named entity is a real-world object, such as persons, locations, organisations, products, etc., that can be denoted with a proper name.



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]**

What might make it more difficult for persons rather than places, and vice versa?

- One common problem, people’s names and places are ambiguous with common nouns

- We can write a list of names of places but cannot write a list of people’s names

- Many locations can have the same name.

What is the IOB trick, in a sequence labelling context? Why is it important?

- IO V.S. IOB

- Apple is looking at buying U.K. startup for $1 billion

Apple	ORG
U.K.	GPE
$1 billion	MONEY

What is Relation Extraction? How is it similar to NER, and how is it different?

- Relation Extraction attempts to find and list the relationships between important events or entities within a document.

- Methods:
  - Rule-based
  - Supervised learning
  - Semi-supervised
  - Distant model
  - Unsupervised

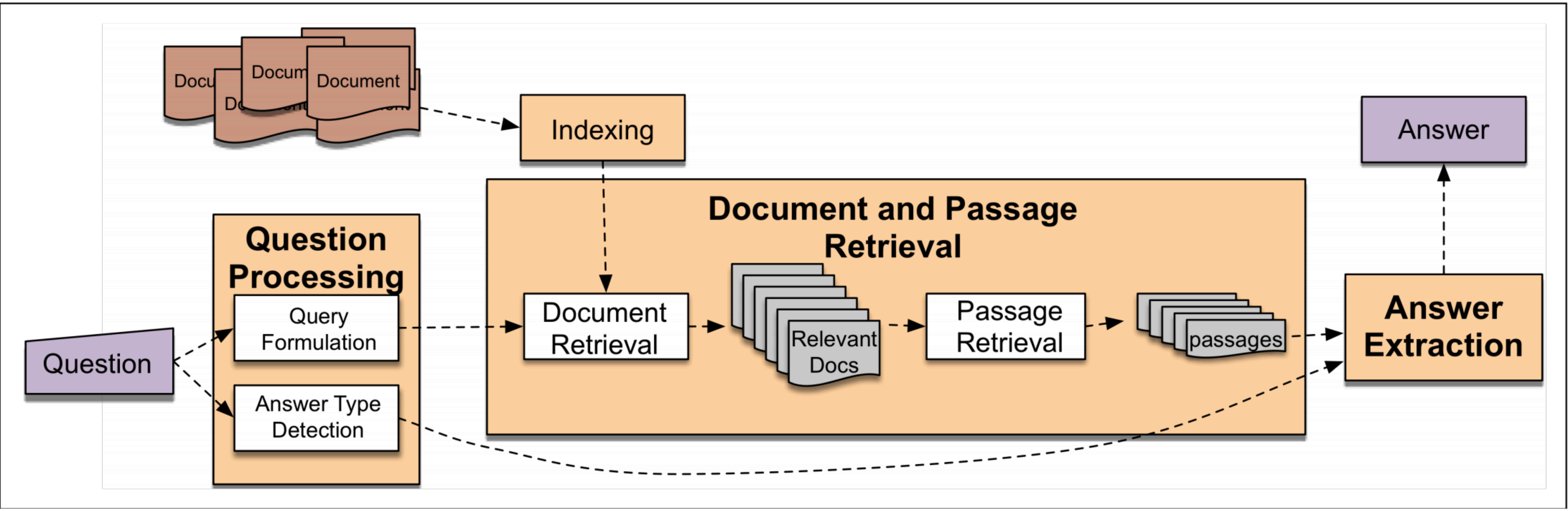
  - E.g. Morgan’s father is Tony. Tony’s wife is pepper.

Supervised relation extraction

- **[ORG American Airlines],** a unit of **[ORG AMR Corp.]**, immediately matched the move, spokesman **[PER Tim Wagner]** said.
- First:
  - (American Airlines, AMR Corp.) -> positive
  - (Tim Wagner, American Airlines) -> positive
  - (Tim Wagner, AMR Corp.) -> negative
- Second:
  - (American Airlines, AMR Corp.) -> subsidiary
  - (Tim Wagner, American Airlines) -> employment

- E.g. Morgan’s father is Tony. Tony’s wife is pepper.

What is Question Answering, and how is it related to Information Retrieval and Information Extraction?



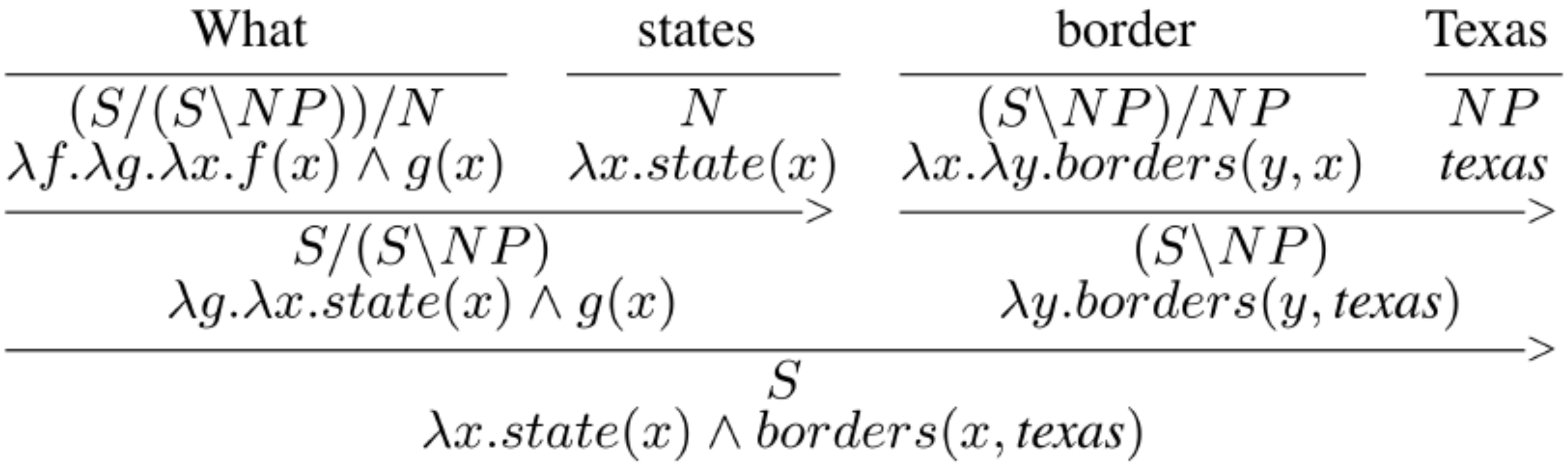
Semantic Parsing

- Based on aligned questions and their logical form, e.g., GeoQuery (Zelle & Mooney 1996)

  What is the capital of the state with the largest population?

  answer(C, (capital(S,C), largest(P, (state(S), population(S,P))))).

- Can model using parsing (Zettlemoyer & Collins 2005) to build compositional logical form


What
states
border
Texas
$(S/(S\backslash NP))/N$
$\lambda f.\lambda g.\lambda x.f(x) \wedge g(x)$
N
$\lambda x.state(x)$
$(S\backslash NP)/NP$
$\lambda x.\lambda y.borders(y,x)$
NP
$texas$
$>$
$S/(S\backslash NP)$
$\lambda g.\lambda x.state(x) \wedge g(x)$
$>$
$(S\backslash NP)$
$\lambda y.borders(y, texas)$
$>$
S
$\lambda x.state(x) \wedge borders(x, texas)$

What might be the main steps for answering a question for a QA system?

- In a Relation Extraction sense:
  - Offline, we process our document collection to generate a list of relations (our knowledge base)

  - When we receive a (textual) query, we transform it into the same structural representation, with some known field(s) and some missing field(s)

  - We examine our knowledge base for facts that match the known fields

  - We rephrase the query as an answer with the missing field(s) filled in from the matching facts from the knowledge base

What might be the main steps for answering a question for a QA system?

- In an Information Retrieval sense:
  – Offline, we process our document collection into a suitable format for IR querying (e.g. inverted index)

  – When we receive a (textual) query, we remove irrelevant terms, and (possibly) expand the query with related terms

  – We select the best document(s) from the collection based on our querying model (e.g. TF-IDF with cosine similarity)

  – We identify one or more snippets from the best document(s) that match the query terms, to form an answer