Wrap-Up: Current Topics in Deep Processing Methods

LING 571 — Deep Processing Methods in NLP

December 3rd, 2018

Ryan Georgi

Coreference Resolution Humor

Coreference Resolution Humor pt. 2

A young artist exhibits his work for the first time and a well known art critic is in attendance.

The critic says to the young artist, "would you like my opinion on your work?"

"Yes, " says the artist.

"It's worthless," says the critic

The artist replies, "I know, but tell me anyway."

Coreference Resolution Humor pt. 2

A young artist exhibits his work for the first time and a well known art critic is in attendance.

The critic says to the young artist, "would you like **my opinion** on **your work**?"

"Yes, " says the artist.

"**It's** worthless," says the critic

The artist replies, "I know, but tell me anyway."

End-to-End Neural Coreference Resolution

Lee e. al, 2017

End-to-End Neural Coreference Resolution

Lee e. al, 2017

- Begin with dataset with gold mention clusters (aka chains)

"General Electric said the Postal Service contacted the company."

End-to-End Neural Coreference Resolution

Lee e. al, 2017

- Can think of the coref problem as finding the maximally likely distribution:

$$P(y_1, \ldots, y_N \mid D) = \prod_{i=1}^{N} \frac{\exp\left(\boxed{s}(i, y_i)\right)}{\sum_{y' \in y(i)} \exp\left(\boxed{s}(i, y')\right)}$$

Where

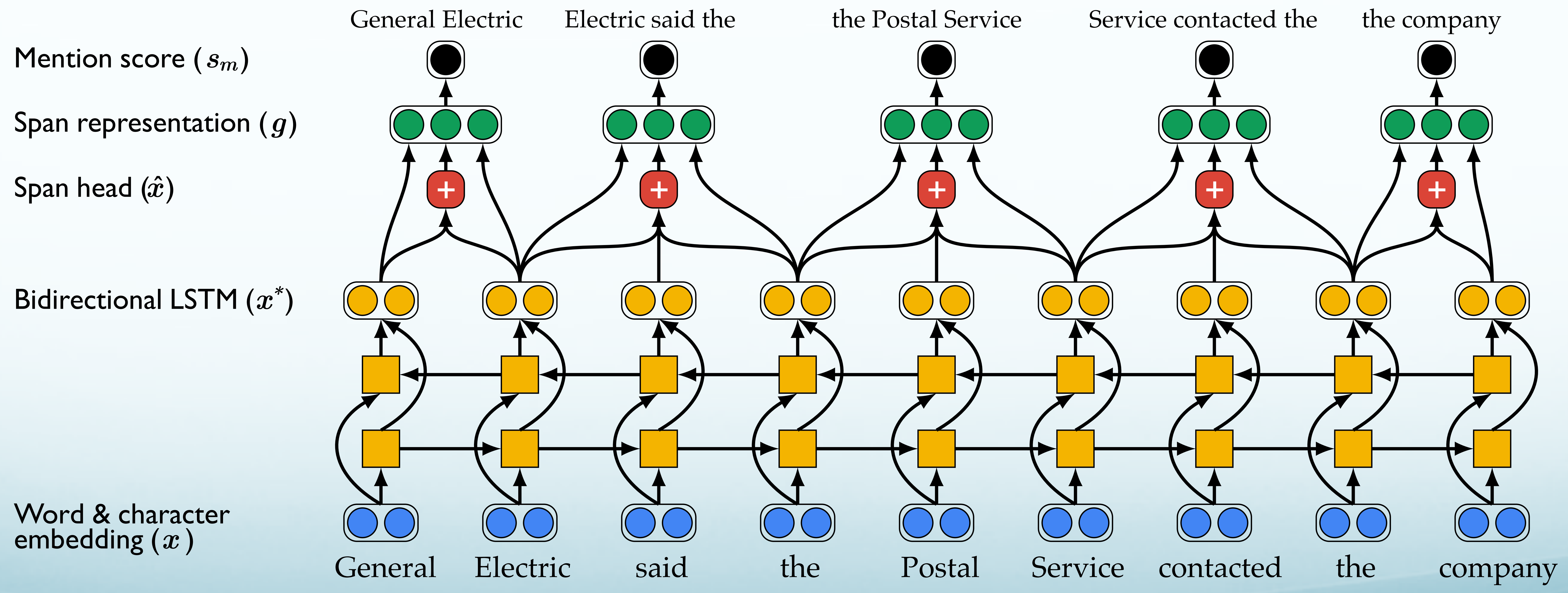
$$\boxed{s(i,j)} = \begin{cases} 0 & j = \epsilon \\ \boxed{s_m}(i) + \boxed{s_m}(j) + \boxed{s_a}(i,j) & j \neq \epsilon \end{cases}$$

Span Score Mention Score Antecedent Score

End-to-End Neural Coreference Resolution

Lee e. al, 2017

- **Step 1** — Train model to identify spans based on gold span labels
  - Use bi-LSTMs to model sequential information preceding/following/within spans
  - Include “headedness” of span with an ***attention*** mechanism



End-to-End Neural Coreference Resolution

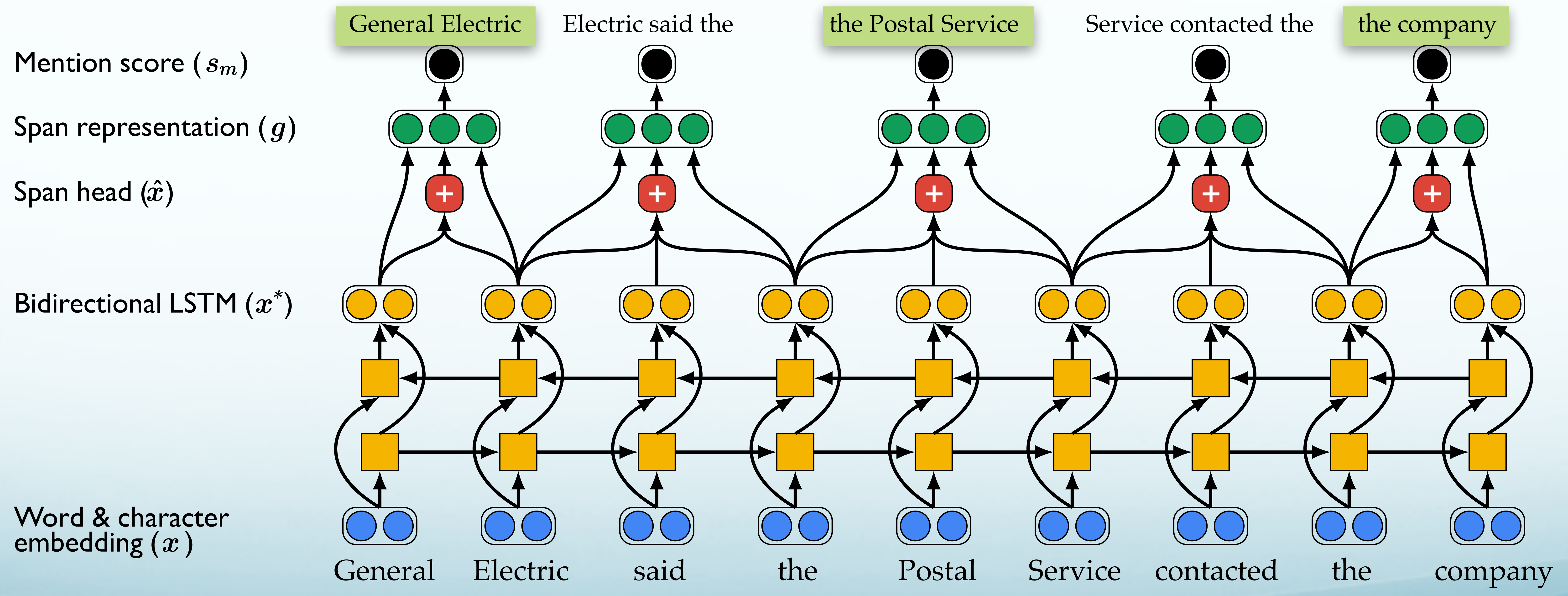
Lee e. al, 2017

- *Attention* can be visualized by heatmap over spans:
  - (The flight attendants) have until 6:00 today to ratify labor concessions. (The pilots') union and ground crew did so yesterday.
  - (Prince Charles and his new wife Camilla) have jumped across the pond and are touring the United States making (their) first stop today in New York. It's Charles' first opportunity to showcase his new wife, but few Americans seem to care. Here's Jeanie Mowth. What a difference two decades make. (Charles and Diana) visited a JC Penney's on the prince's last official US tour. Twenty years later, here's the prince with his new wife.

End-to-End Neural Coreference Resolution

Lee e. al, 2017

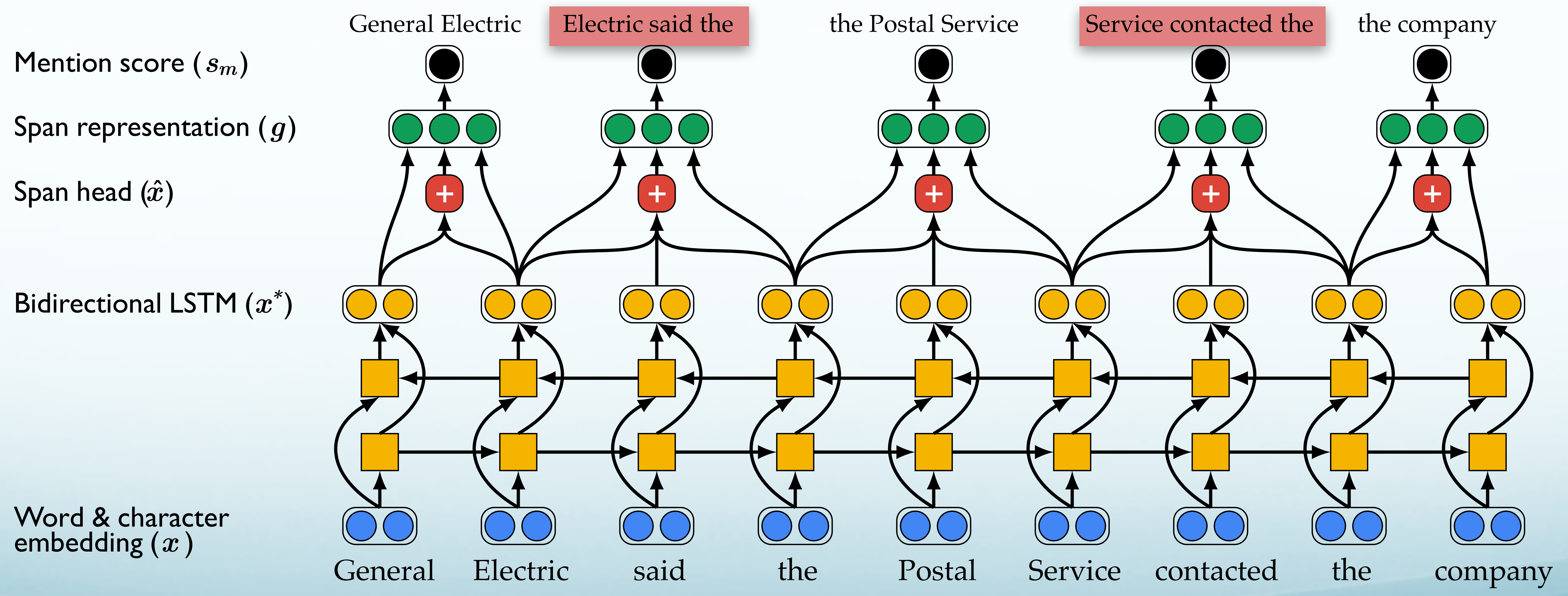
- **These** are valid gold mentions (network gets "reward" for getting these right)



End-to-End Neural Coreference Resolution

[Lee e. al, 2017]

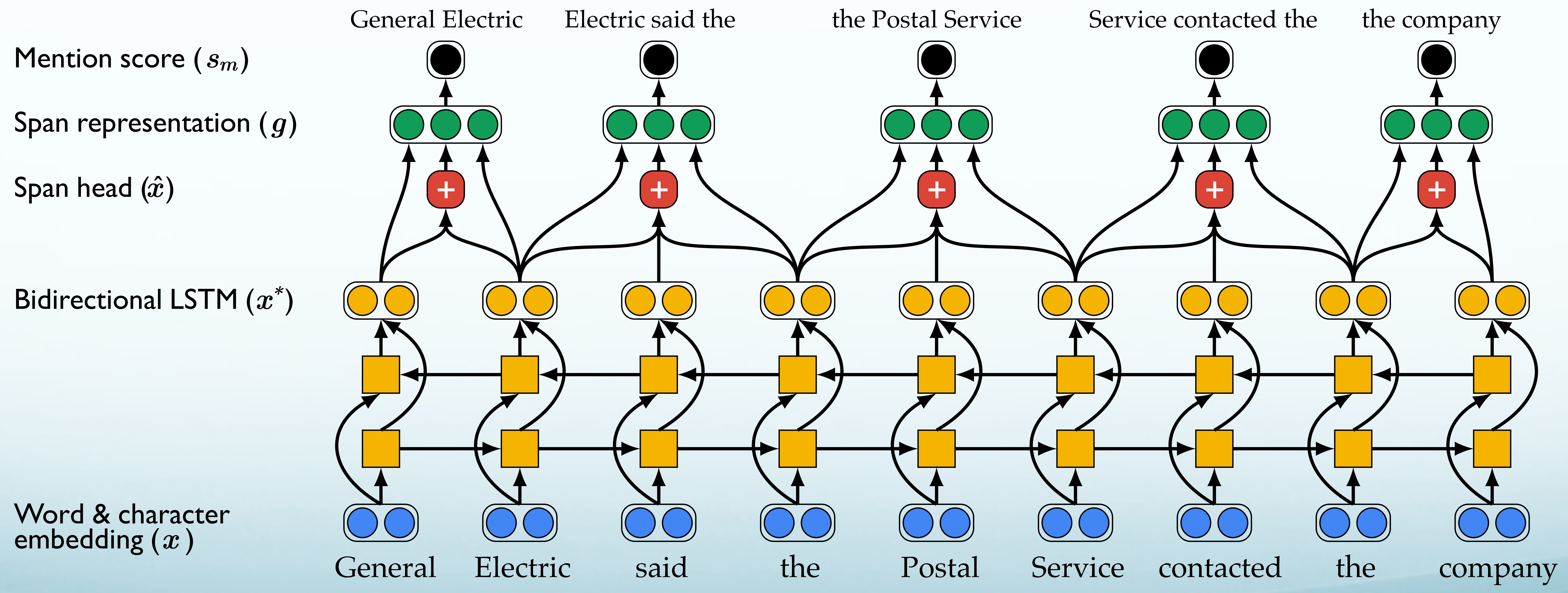
- **These** are invalid mentions (network accumulates error if these are selected)



End-to-End Neural Coreference Resolution

Lee e. al, 2017

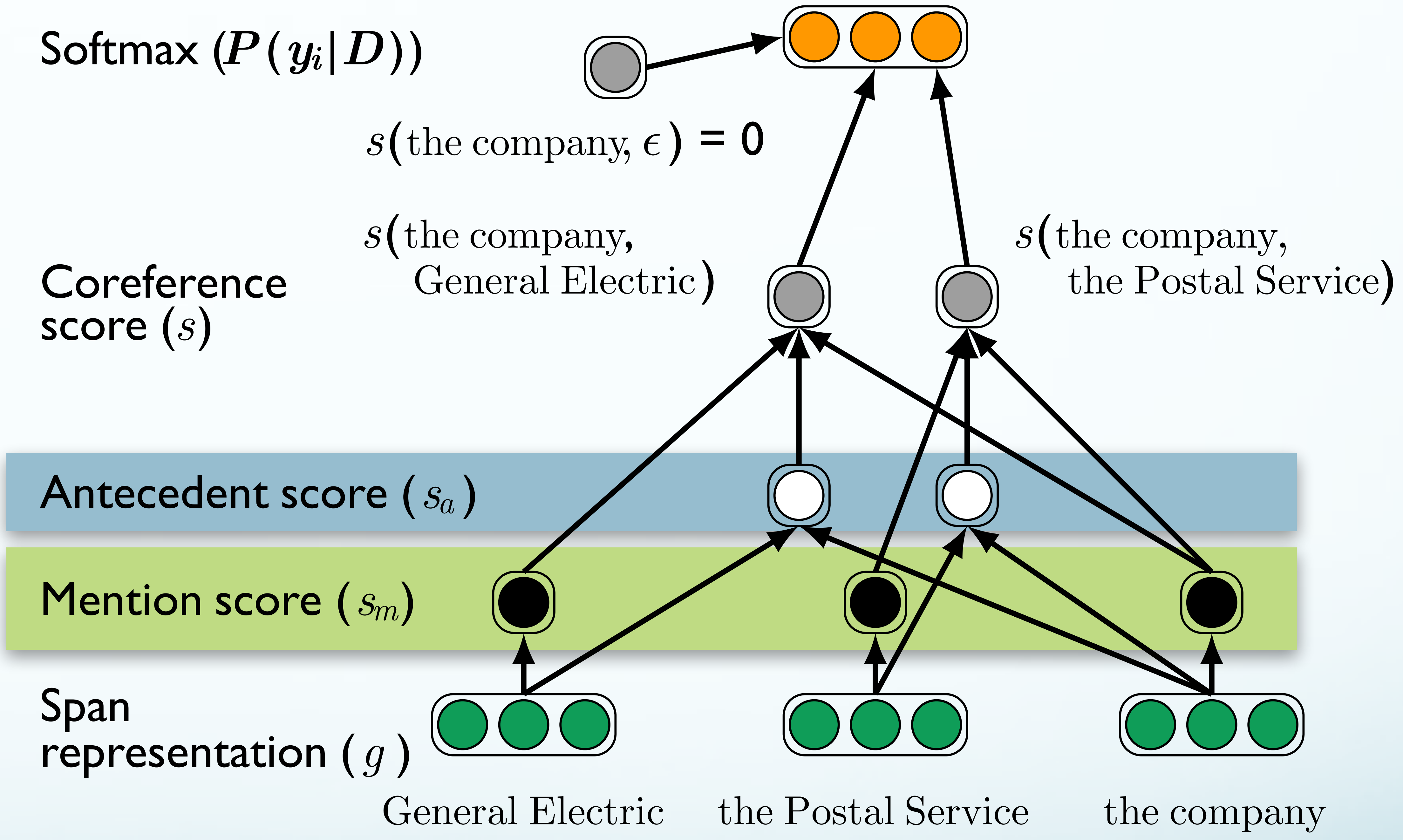
- Network thus learns to identify features from (embeddings → sequence) + head
  - As more or less likely to identify a span of words as a mention



End-to-End Neural Coreference Resolution

Lee e. al, 2017

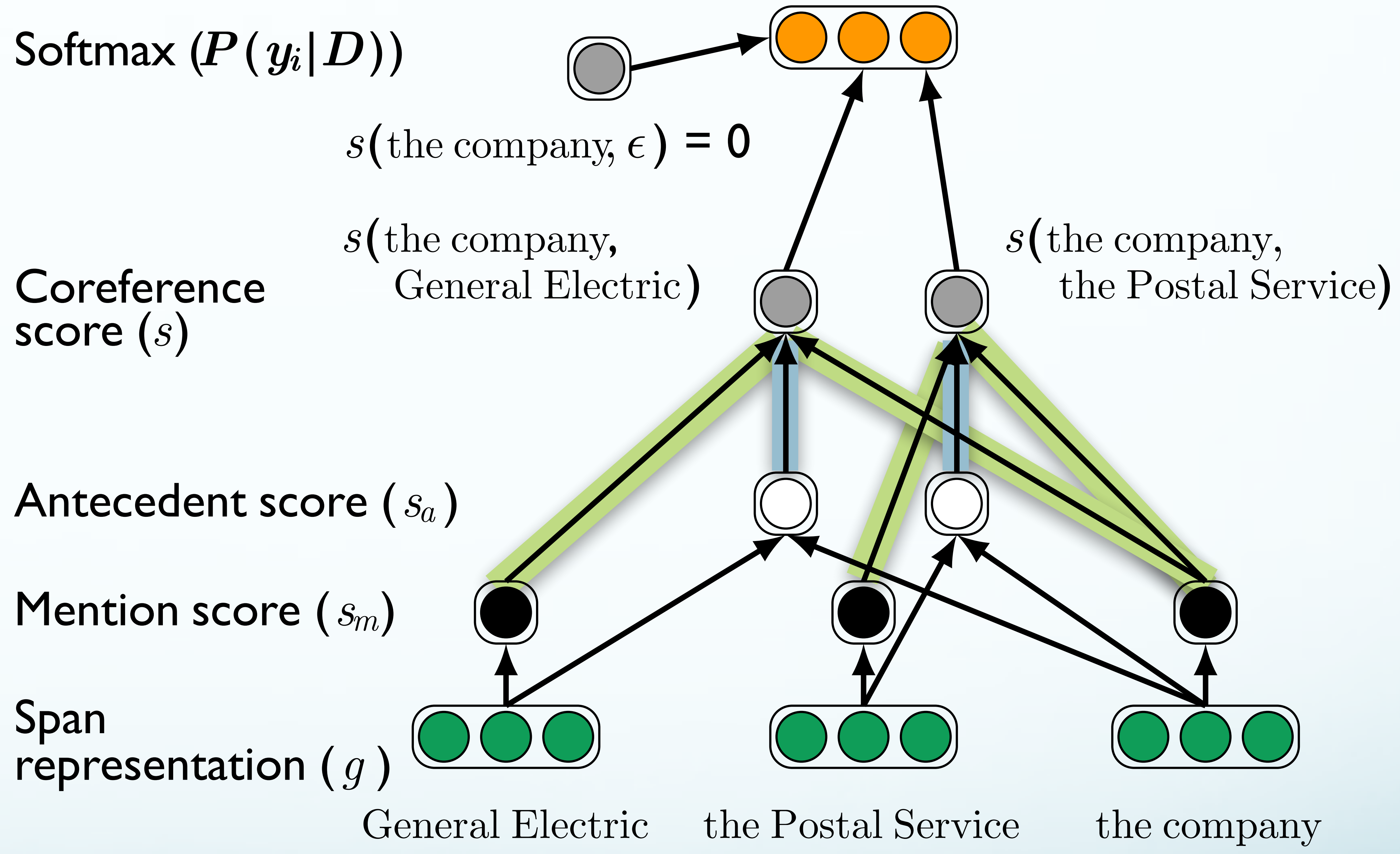
- **Step 2** — Learn Coref Clusters
- Mention Scores
  - Likelihood a given span is a mention
  - Unary over spans
- Antecedent scores
  - Likelihood another mention is an antecedent
  - Pairwise between spans



End-to-End Neural Coreference Resolution

Lee e. al, 2017

- The coref score is a combination of:
  - antecedent scores
  - mention scores



End-to-End Neural Coreference Resolution

[Lee e. al, 2017]

- Other info:
  - Also implement pruning to avoid dealing with *all* spans
  - Also encode metadata, such as speaker and genre in mention representation

End-to-End Neural Coreference Resolution

Lee e. al, 2017

- Data:
  - CoNLL-2012 Shared Task (Coref on OntoNotes)
  - 2802 training docs
  - 343 development docs
  - 348 test docs
  - 454 words/doc average

End-to-End Neural Coreference Resolution

Lee e. al, 2017

- Positive:
  - State-of-the-art on CoNLL-2012 Test Data
- Errors:
  - Word embeddings tend to conflate paraphrasing with relatedness
    - e.g. (The flight attendants) have until 6:00 today to ratify labor concessions. (The pilots') union and ground crew did so yesterday.
  - (Prince Charles and his new wife Camilla) have jumped across the pond … (Charles and Diana) visited a JC Penney's…

Neural Sequence Learning Models for Word Sense Disambiguation

Raganato et. al (2017b)

Neural Sequence Learning Models for WSD

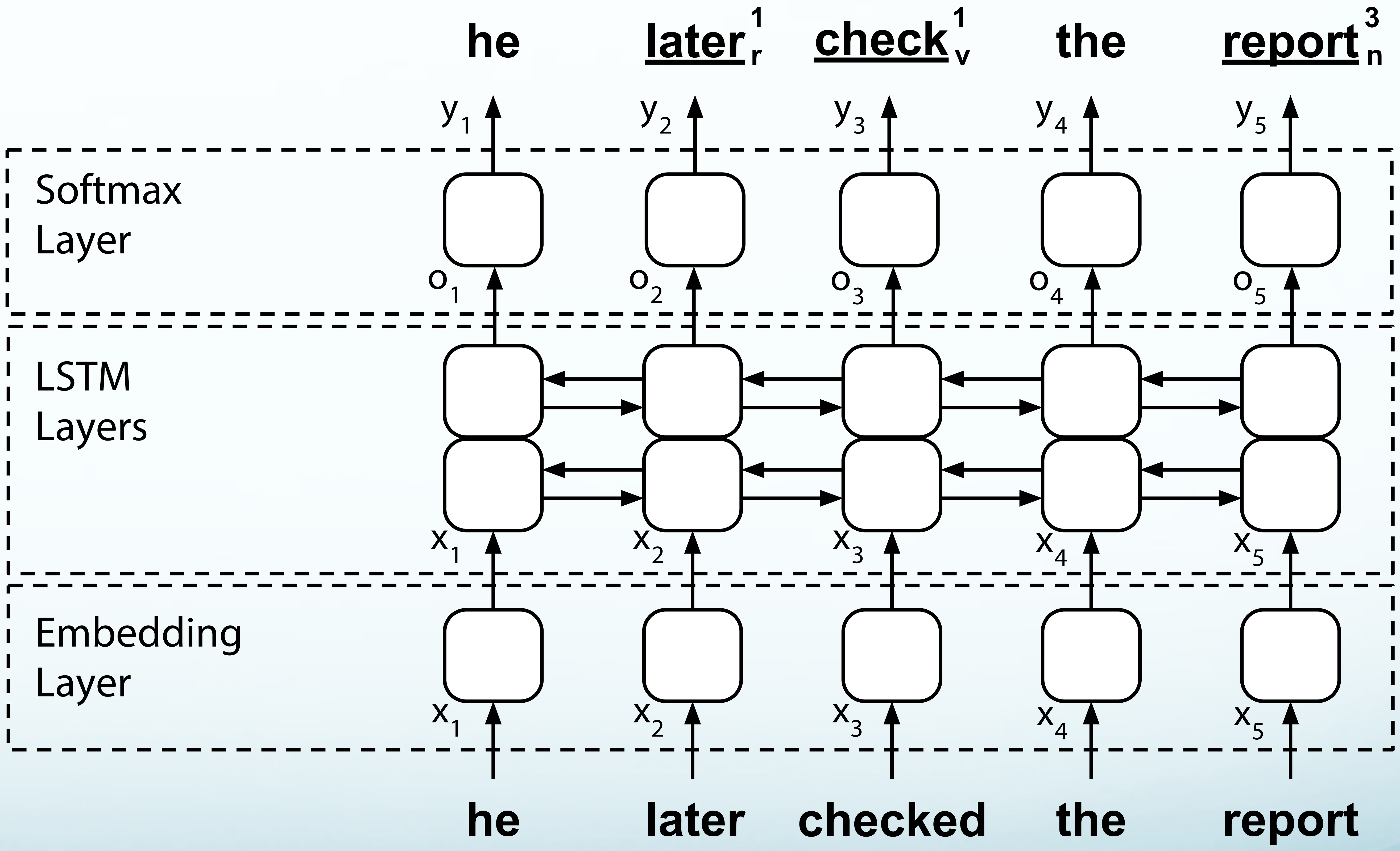
Raganato et. al (2017b)

- Authors propose several models for encoding words and senses
  - **bi-LSTM**
  - **bi-LSTM + Attention**
  - **Sequence to Sequence**
- All approaches are encoding sequential information
- All approaches use sense-tagged corpus

Neural Sequence Learning Models for WSD

Raganato et. al (2017b)

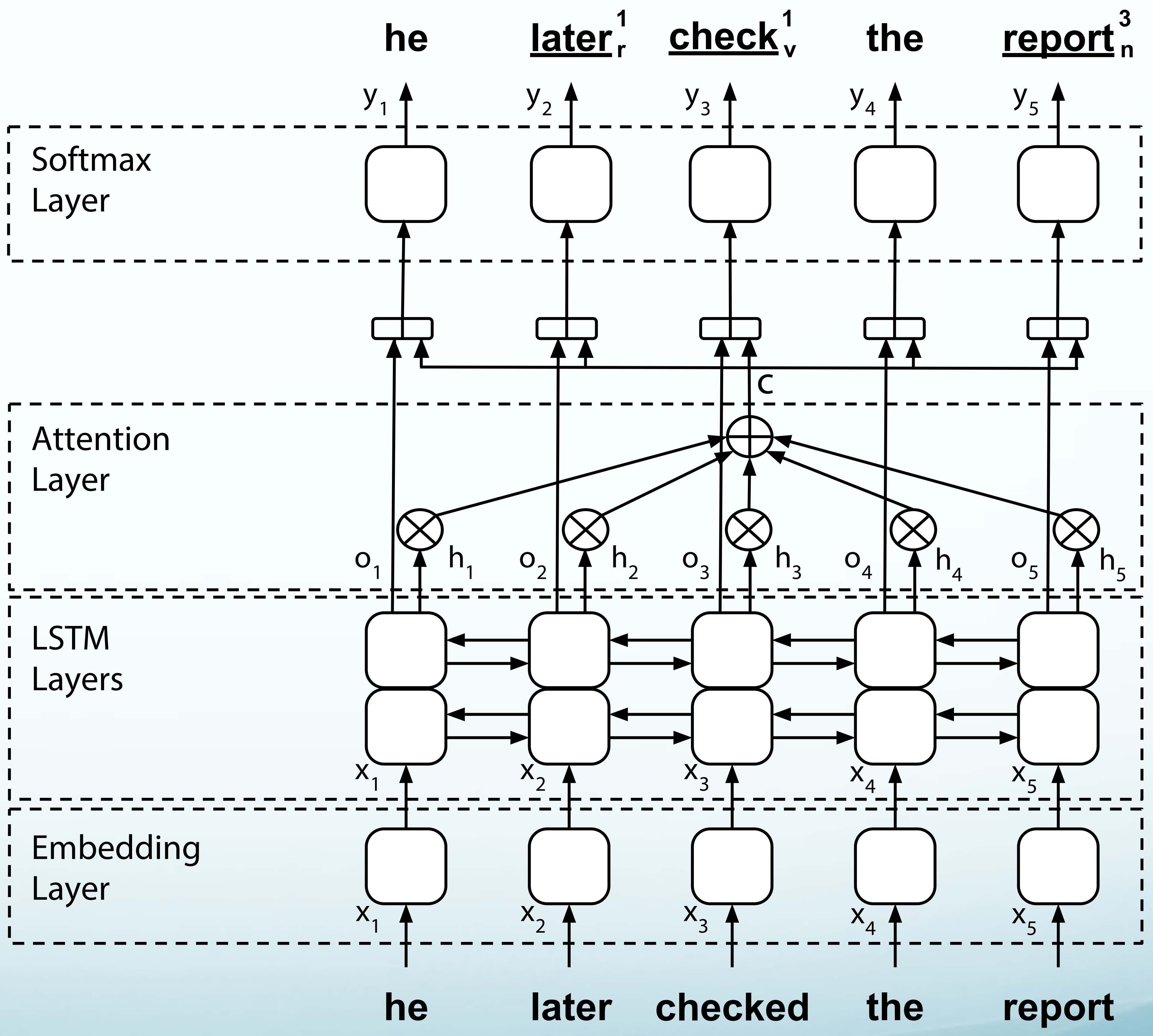
- **bi-LSTM**
  - Learn to label proper sense given word embedding and context (LSTM)



Neural Sequence Learning Models for WSD

Raganato et. al (2017b)

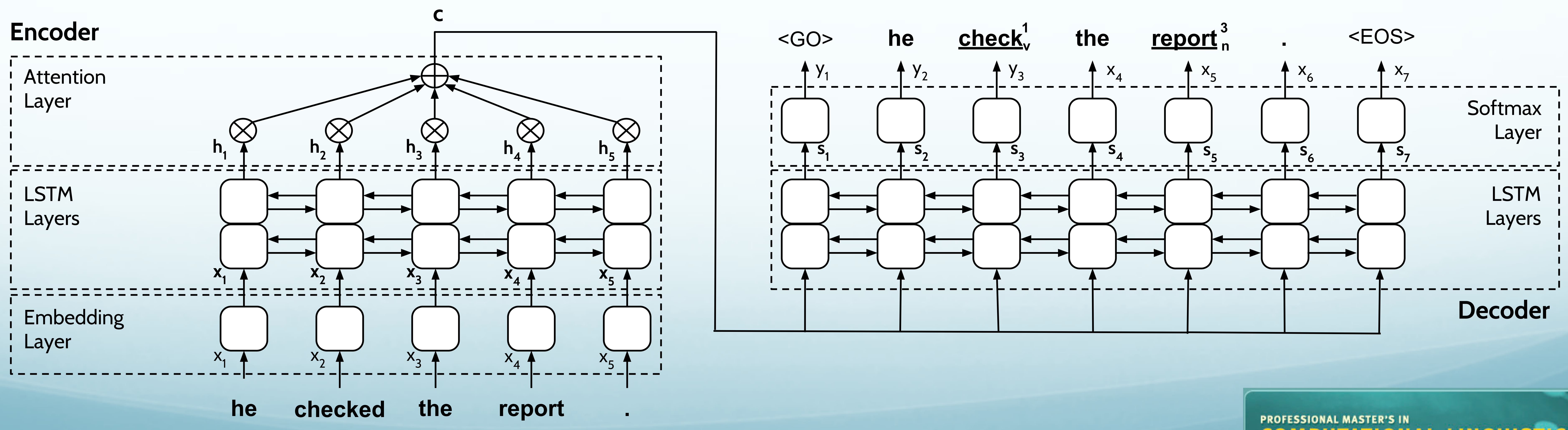
- **bi-LSTM + Attention**
  - Attention layer adds sentence-level representation *c* to guide the labels generate at each sequence time step by focusing on what part of the sentence may be relevant
    - (e.g. with *wicket* in focus, *match* might be influenced toward the game sense, rather than firestarter)



Neural Sequence Learning Models for WSD

[Raganato et. al (2017b)](#)

- **seq2seq**
  - Two-step task:
    - Memorization — Model is trained to replicate input token-by-token
    - Disambiguation — Model learns to replace surface forms with appropriate senses



Neural Sequence Learning Models for WSD

Raganato et. al (2017b)

- Also try models that jointly learn WSD and:
  - course semantic labels
    - e.g. $noun.location$, $verb.motion$
  - POS tags
  - Both

Neural Sequence Learning Models for WSD

Raganato et. al (2017b)

- Data:
  - Use SemCor 3.0 for training/evaluating word senses

Neural Sequence Learning Models for WSD

Raganato et. al (2017b)

- Results:

	Dev	Test Datasets				Concatenation of All Test Datasets				
	SE07	SE2	SE3	SE13	SE15	Nouns	Verbs	Adj.	Adv.	All
BLSTM	61.8	71.4	68.8	65.6	69.2	70.2	56.3	75.2	**84.4**	68.9
BLSTM + att.	62.4	71.4	**70.2**	66.4	70.8	71.0	**58.4**	75.2	83.5	69.7
BLSTM + att. +LEX	63.7	**72.0**	69.4	66.4	**72.4**	**71.6**	57.1	**75.6**	83.2	**69.9**
BLSTM + att. +LEX +POS	**64.8**	**72.0**	69.1	**66.9**	71.5	71.5	57.5	75.0	83.8	**69.9**
Seq2Seq	60.9	68.5	67.9	65.3	67.0	68.7	54.5	74.0	81.2	67.3
Seq2Seq + att.	62.9	69.9	69.6	65.6	67.7	69.5	57.2	74.5	81.8	68.4
Seq2Seq + att. +LEX	64.6	70.6	67.8	66.5	68.7	70.4	55.7	73.3	82.9	68.5
Seq2Seq + att. +LEX +POS	63.1	70.1	68.5	66.5	69.2	70.1	55.2	75.1	84.4	68.6
IMS	61.3	70.9	69.3	65.3	69.5	70.5	55.8	75.6	82.9	68.9
IMS+emb	**62.6**	**72.2**	**70.4**	65.9	71.5	**71.9**	56.6	**75.9**	**84.7**	**70.1**
Context2Vec	61.3	71.8	69.1	65.6	**71.9**	71.2	**57.4**	75.2	82.7	69.6
Lesk_{ext} +emb	56.7	63.0	63.7	66.2	64.6	70.0	51.1	51.7	80.6	64.2
UKB_{gloss} w2w	42.9	63.5	55.4	62.9	63.3	64.9	41.4	69.5	69.7	61.1
Babelfy	51.6	67.0	63.5	**66.4**	70.3	68.9	50.7	73.2	79.8	66.4
MFS	54.5	65.6	66.0	63.8	67.1	67.7	49.8	73.1	80.5	65.5

Neural Sequence Learning Models for WSD

Raganato et. al (2017b)

- Analysis:
  - Comparable to other supervised systems
  - Adding course-grained lexical tags appears to help
  - POS did not seem to help
- ***None of these systems substantially better than using the Most Frequent Sense***

Deep Contextualized Word Representations

Peters et. al (2018)

Deep Contextualized Word Representations

Peters et. al (2018)

- Prior vector-space embeddings have typically been derived:
  - Context-independent distributions (CBOW)
  - CNNs over characters

Deep Contextualized Word Representations

Peters et. al (2018)

- NAACL 2018 Best Paper Award
- **E**mbeddings from **L**anguage **Mo**dels (ELMo)
- Rather than treat embeddings as bag of words
  - Create embeddings by using sequential modeling (bi-LSTM)

Deep Contextualized Word Representations

Peters et. al (2018)

- Comparison to GloVe:

	Source	Nearest Neighbors
GloVe	play	playing, game, games, played, players, plays, player, Play, football, multiplayer
biLM	Chico Ruiz made a spectacular **play** on Alusik's grounder…	Kieffer, the only junior in the group, was commended for his ability to hit in the clutch, as well as his all-round excellent **play**.
	Olivia De Havilland signed to do a Broadway **play** for Garson…	…they were actors who had been handed fat roles in a successful **play**, and had talent enough to fill the roles competently, with nice understatement.

Deep Contextualized Word Representations

Peters et. al (2018)

- Intrinsic evaluation via WSD:

Model	F_1
WordNet 1st Sense Baseline	65.9
Raganato et al. (2017a)	69.9
Iacobacci et al. (2016)	70.1
CoVe, First Layer	59.4
CoVe, Second Layer	64.7
biLM, First layer	67.4
biLM, Second layer	69.0

Deep Contextualized Word Representations

Peters et. al (2018)

- Used in place of other embeddings on multiple tasks:

Task	Previous SOTA		Our baseline	ELMo + baseline	Increase (absolute/relative)
SQuAD	Liu et al. (2017)	84.4	81.1	85.8	4.7 / 24.9%
SNLI	Chen et al. (2017)	88.6	88.0	88.7 ± 0.17	0.7 / 5.8%
SRL	He et al. (2017)	81.7	81.4	84.6	3.2 / 17.2%
Coref	Lee et al. (2017)	67.2	67.2	70.4	3.2 / 9.8%
NER	Peters et al. (2017)	91.93 ± 0.19	90.15	92.22 ± 0.10	2.06 / 21%
SST-5	McCann et al. (2017)	53.7	51.4	54.7 ± 0.5	3.3 / 6.8%

SQuAD = Stanford Question Answering Dataset
SNLI = Stanford Natural Language Inference Corpus
SST-5 = Stanford Sentiment Treebank

Question-Answering:

A Case Study in Shallow vs. Deep Methods

Question Answering: The Problem

- Grew out of information retrieval community
- Document retrieval is great, but…
  - Sometimes you don't just want a ranked list of documents.
  - Sometimes you want an answer to a question
    - Short answer, possibly with supporting context
- People ask questions on the web
  - *Which English translation of the Bible is used in official Catholic liturgies?*
  - *Who invented surf music?*
  - *What are the seven wonders of the world?*
  - These account for 12–15% of web log queries

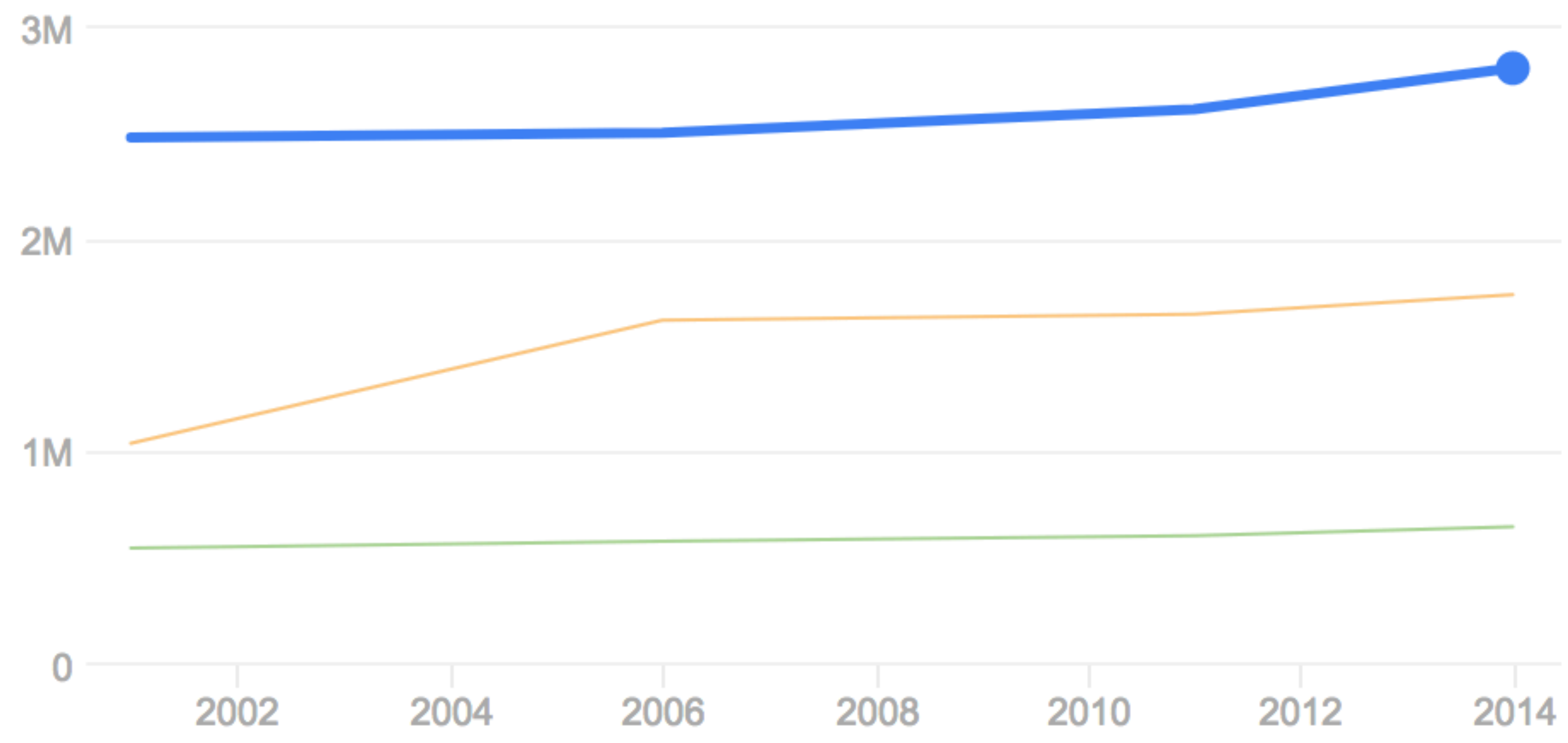
Search Engines and Questions

- What do search engines do with questions?
  - Increasingly, try to answer questions
  - Especially for Wikipedia infobox types of info
  - Backoff to keyword search
- How well does this work?

What Canadian city has the largest population?

Toronto / Population

2.809 million (2014)



- Toronto 2.809 million
- Montreal 1.741 million
- Vancouver 647,540

Explore more

Sources include: UNdata   *Feedback*

People also ask

What are the 3 largest cities in Canada by population?

What are the 5 major cities in Canada?

Toronto

City in Ontario, Canada

Toronto, the capital of the province of Ontario, is a major Canadian city along Lake Ontario's northwestern shore. It's a dynamic metropolis with a core of soaring skyscrapers, all dwarfed by the iconic, free-standing CN Tower. Toronto also has many green spaces, from the orderly oval of Queen's Park to 400-acre High Park and its trails, sports facilities and zoo.

Population elsewhere

Canada	35.54 million (2014)
New York City	8.472 million (2014)
Chicago	2.719 million (2014)

Sources include: World Bank, United States Census Bureau   *Feedback*

What is the total population of the ten largest capitals in the US?

- Rank 1 snippet:
  - As of 2013, 61,669,629 citizens lived in ***America's*** 100 ***largest cities***, which was 19.48 percent of the nation's ***total population***.
  - See the top 50 ***U.S. cities by population*** and rank. … The table below lists the *largest* 50 *cities in the*
  - The table below lists the *largest* 10 *cities in the United States*…

Search Engines and QA

- Search for exact question string
  - "Do I need a visa to go to Japan?"
    - Result: Exact match on Yahoo! Answers
    - Find "Best Answer" and return following chunk
- Works great… if the question matches exactly
  - Many websites are building archives
  - What happens if it doesn't match?
    - "Question mining" tries to learn paraphrases of questions to get answers.

Perspectives on QA

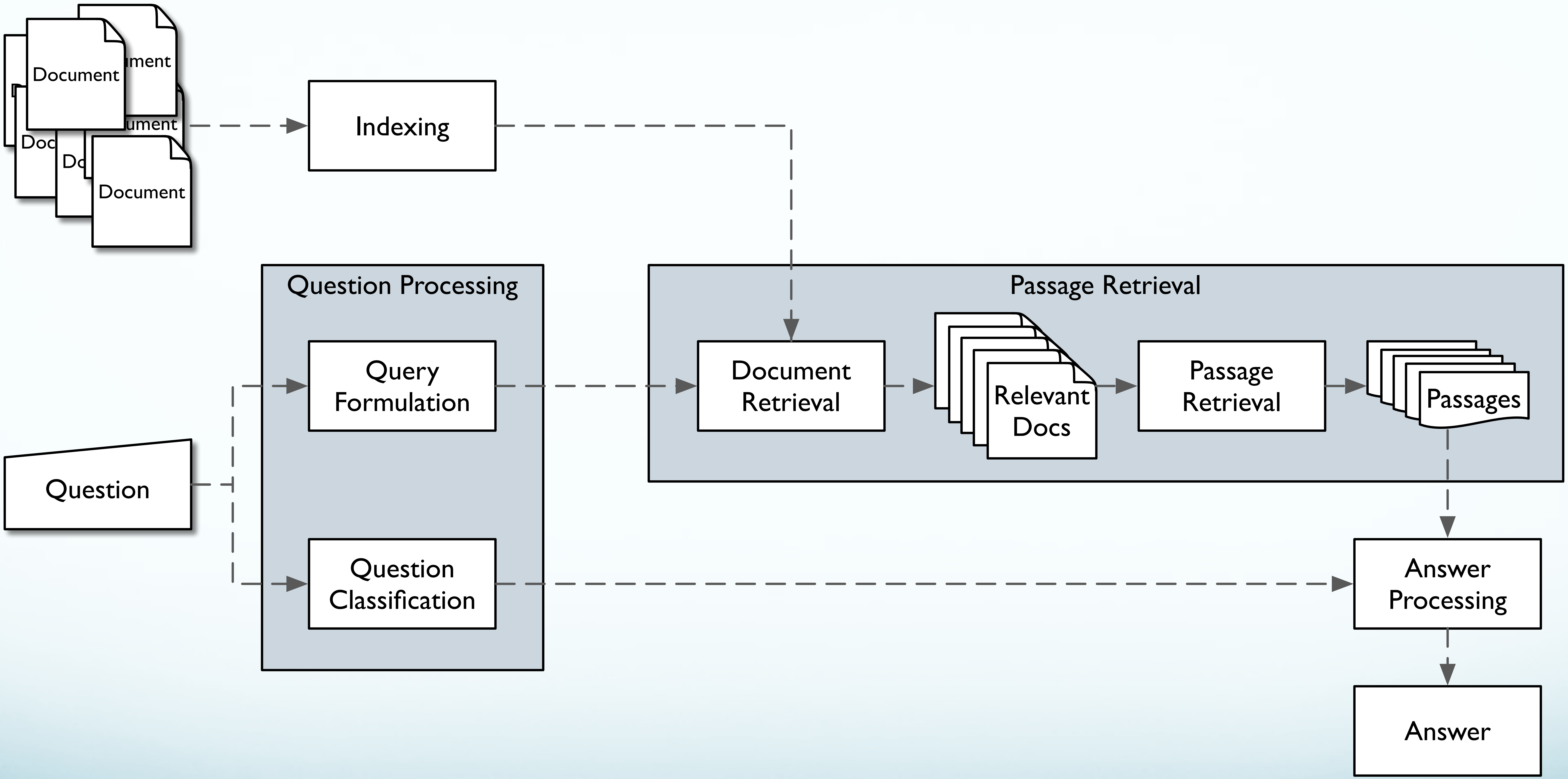
- TREC QA track (~2000—)
  - Initially pure factoid questions, with fixed length answers
    - Based on large collection of fixed documents (news)
    - Increasing complexity: definitions, biographical info, etc
      - Single response
- Reading comprehension (Hirschman et al, 1999—)
  - Think SAT/GRE
    - Short text or article (usually middle school level)
    - Answer questions based on text
  - Also, "Machine Reading"
- And, of course, Jeopardy! and Watson

Question Answering (*a la* TREC)

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

Basic Strategy

- Given an indexed document collection…
- …and a question…
- …execute the following steps:
  - Query Formulation
  - Question Classification
  - Passage Retrieval
  - Answer Processing
  - Evaluation


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

Query Processing: Query Reformulation

- Query reformulation
  - Convert question to suitable form for IR
  - e.g. "Stop Structure" removal:
    - Delete function words, q-words, even low content verbs

Query Processing: Question Classification

- Answer type recognition:
  - *Who…* → Person
  - *What Canadian City…* → City
  - *What is surf music…* → Definition
- Train classifiers to recognize expected answer type
  - Using POS, NE, words, synsets, hyper/hyponyms

HUMAN	
description	Who was Confucius?
group	What are the major companies that are part of Dow Jones?
ind	Who was the first Russian astronaut to do a spacewalk?
title	What was Queen Victoria's title regarding India?
LOCATION	
city	What's the oldest capital city in the Americas?
country	What country borders the most others?
mountain	What is the highest peak in Africa?
other	What river runs through Liverpool?
state	What states do not have state income tax?
NUMERIC	
code	What is the telephone number for the University of Colorado?
count	About how many soldiers died in World War II?
date	What is the date of Boxing Day?
distance	How long was Mao's 1930s Long March?
money	How much did a McDonald's hamburger cost in 1963?
order	Where does Shanghai rank among world cities in population?
other	What is the population of Mexico?
period	What was the average life expectancy during the Stone Age?
percent	What fraction of a beaver's life is spent swimming?
speed	What is the speed of the Mississippi River?
temp	How fast must a spacecraft travel to escape Earth's gravity?
size	What is the size of Argentina?
weight	How many pounds are there in a stone?

Passage Retrieval

- Why not just perform general information retrieval?
  - Documents too big, non-specific for answers
- Identify shorter, focused spans (e.g. sentences)
  - Filter for correct type: answer type classification
  - Rank passages based on a trained classifier
  - Or, for web search, use result snippets

Answer Processing

- Find the specific answer in the passage
- Pattern extraction-based:
  - Include answer types, regular expressions
    - Can use syntactic/dependency/semantic patterns
    - Leverage large knowledge bases

Pattern	**Question**	**Answer**
<AP> such as <QP>	What is autism?	", **developmental disorders** such as autism…"
<QP>, a <AP>	What is a caldera?	"…the Long Valley caldera, a **volcanic crater** 19 miles long…"

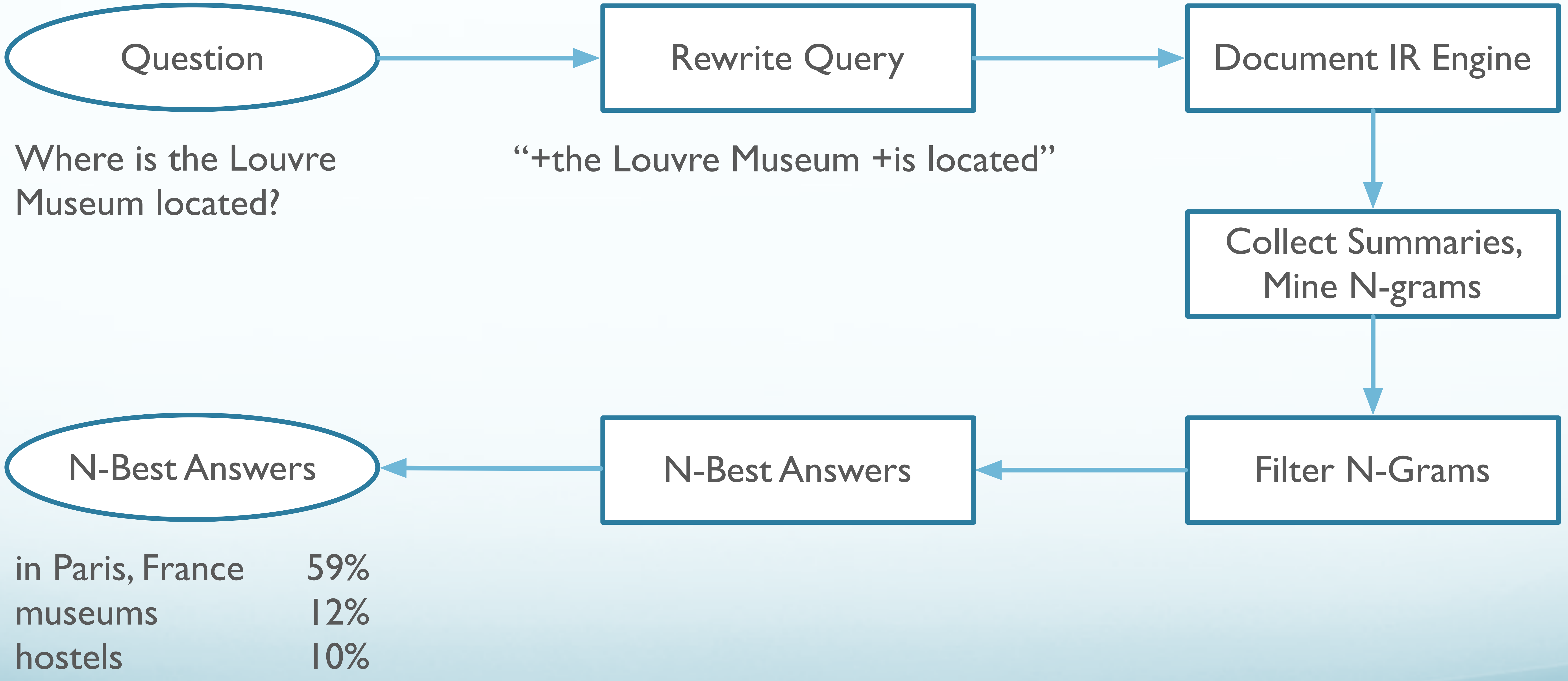
Evaluation

- Classical:
  - Return ranked list of answer candidates
  - Idea: Correct answer higher in list ⇒ higher score
- Measure: Mean Reciprocal Rank (MRR)
  - For each question
    - Get reciprocal of rank of first correct answer
    - e.g. correct answer is 4 ⇒ ¼
    - None correct ⇒ 0
  - Average over all questions

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

AskMSR/Aranea (Lin, Brill)

- Shallow Processing for QA



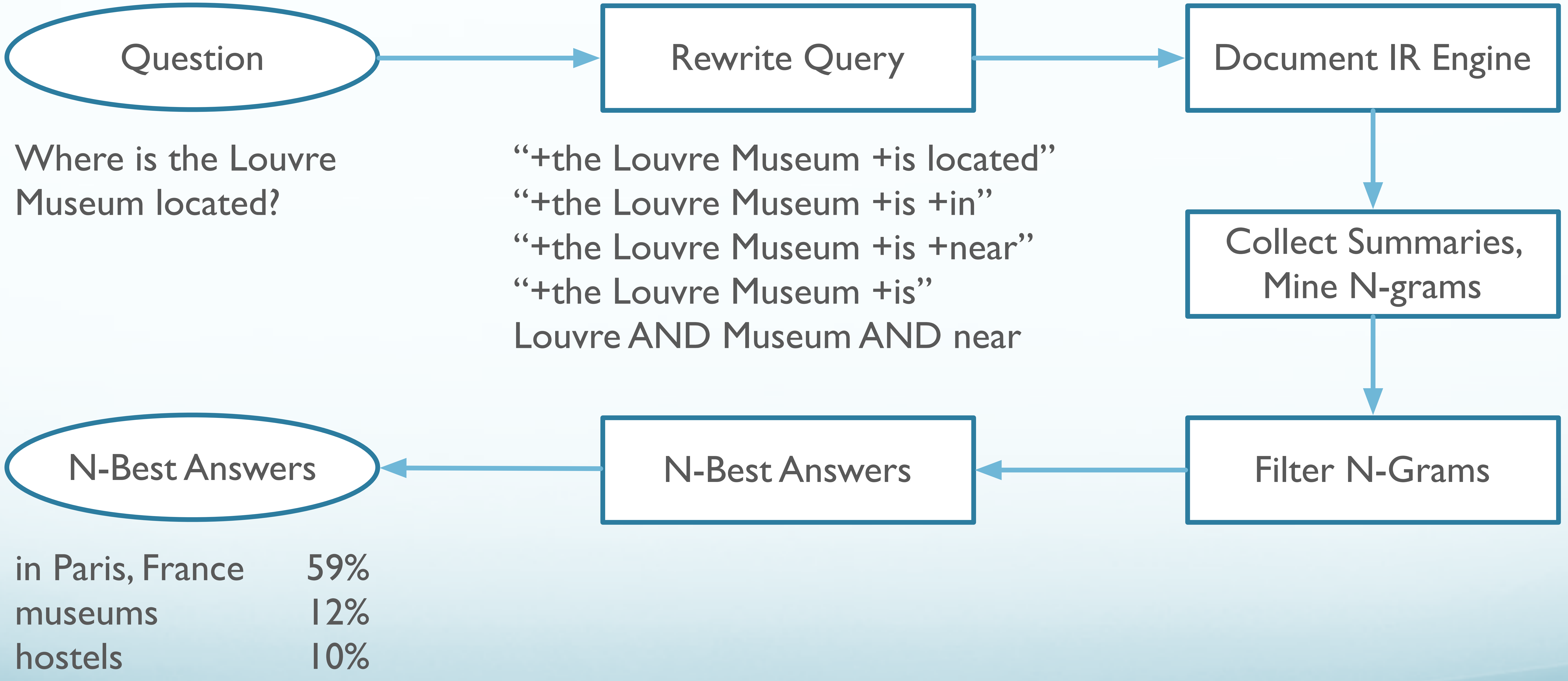
Intuition

- Redundancy is useful!
  - If similar strings appear in many candidate answers, likely to be solution
  - Even if can't find obvious answer strings
- *Q: How many times did Björn Borg win Wimbledon?*
  - Björn Borg *blah blah blah* Wimbledon *blah* 5 *blah*
  - Wimbledon *blah blah blah* Björn Borg *blah* 37 *blah*
  - *blah* Björn Borg *blah blah* 5 *blah blah* Wimbledon
  - 5 *blah blah* Wimbledon *blah blah* Björn Borg
- A: …Probably 5

Retrieval, N-Gram Mining & Filtering

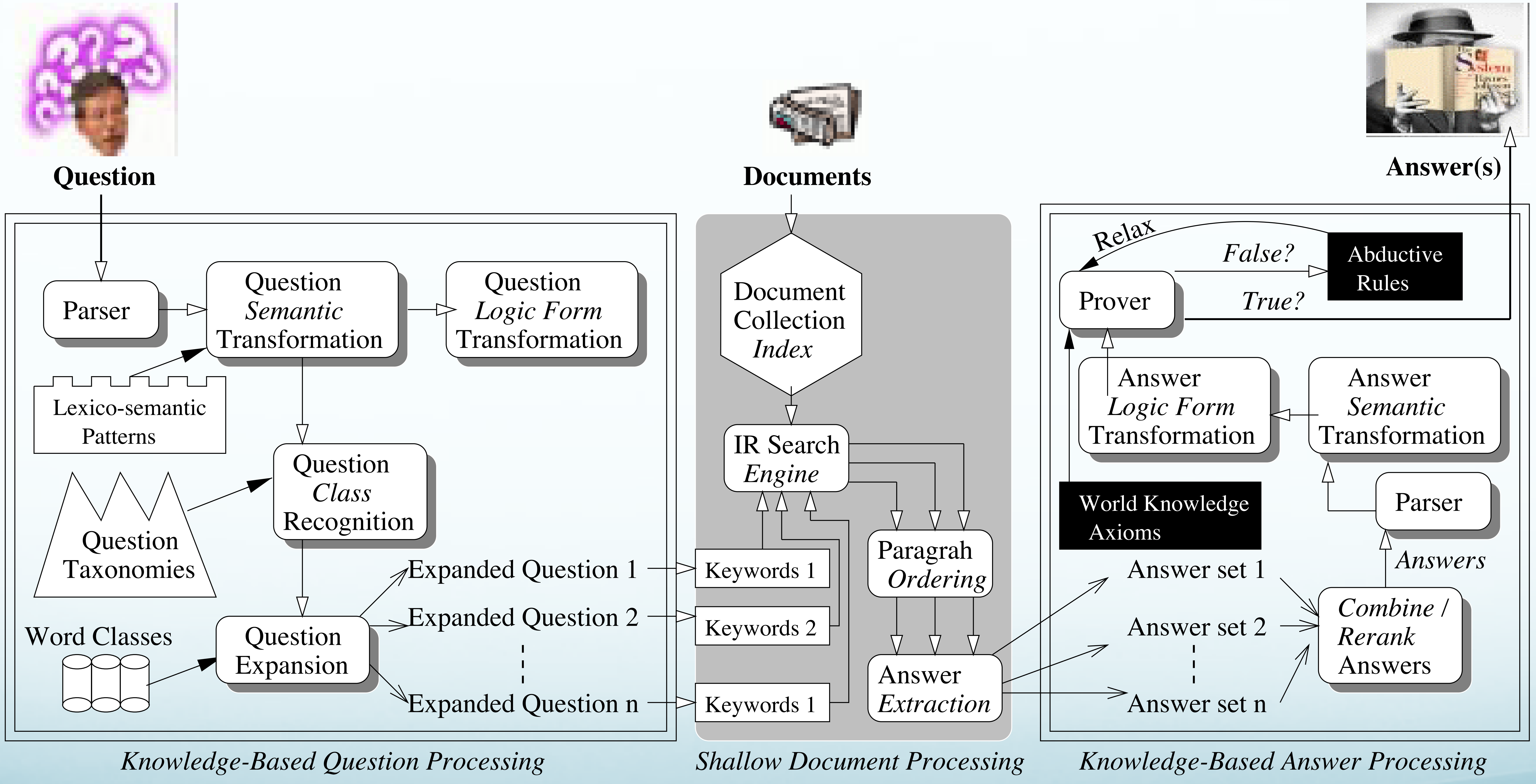
- Run reformulated queries through search engine
  - Collect (lots of) result snippets
  - Collect n-grams from snippets
  - Weight each n-gram summing over occurrences
  - Concatenate n-grams into longer answers
    - e.g. Dickens, Charles Dickens, Mr. Charles

Example Redux



Deep Processing Technique for QA:
LCC PowerAnswer

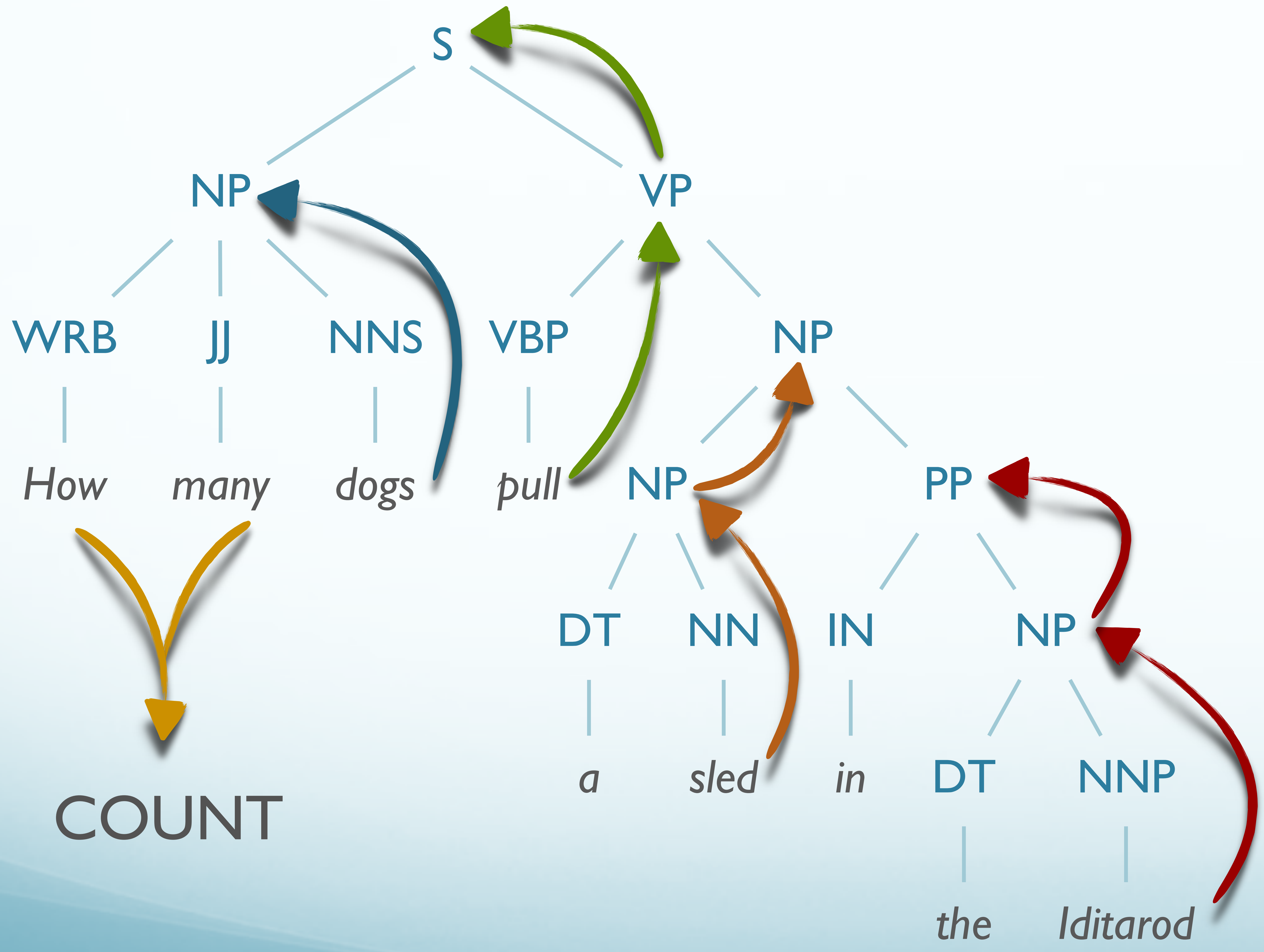
- Experiments with open-domain textual Question Answering, Moldovan, Harabagiu, et al, 2000

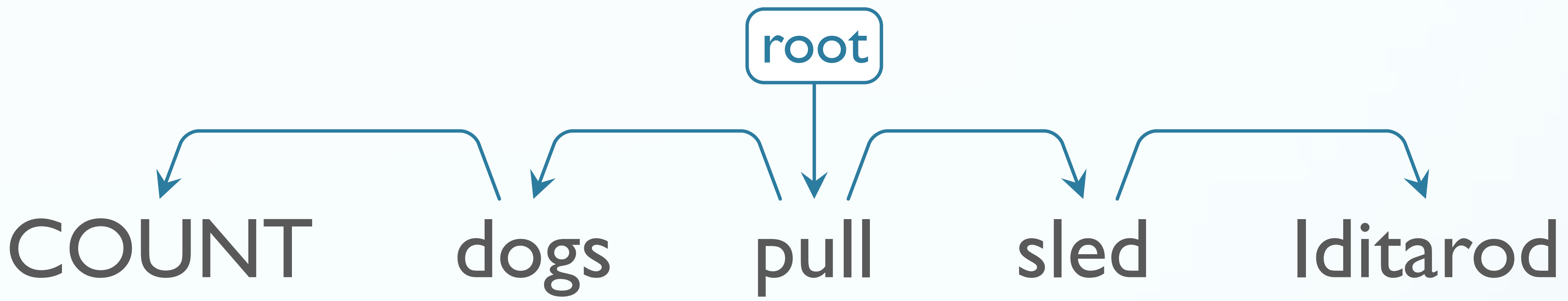


Deep Processing: Query/Answer Formulation

- Preliminary shallow processing:
  - Tokenization, POS tagging, NE recognition, Preprocessing
- Parsing creates syntactic representation:
  - Focused on nouns, verbs, particles
- Coreference resolution links entity references
- Translate to full logical form
  - As close as possible to syntax

Syntax to Logical Form





Deep Processing: Answer Selection

- Lexical Chains:
  - Bridge gap in lexical choice between Question and Answer
    - Improve retrieval and answer selection
  - Create connections via WordNet synsets
    - *Q: When was the internal combustion engine invented?*
    - *A: The first internal-combustion engine was built in 1867.*
    - *invent → create_mentally → create → build*
- Perform abductive reasoning
  - Try to justify answer given question → 30% improvement in accuracy!

A Victory for Deep Processing:

TREC 2002 QA Track

Run Tag	Confidence weighted Score	Correct Answers #	Correct Answers %	Number Inexact	NIL Accuracy Prec	NIL Accuracy Recall
LCCmain2002	**0.856**	**415**	**83.0**	**8**	**0.578**	**0.804**
exactanswer	0.691	271	54.2	12	0.222	0.848
pris2002	0.610	290	58.0	17	0.241	0.891
IRST02D1	0.589	192	38.4	17	0.167	0.217
IBMPQSQACYC	0.588	179	35.8	9	0.196	0.630
uwmtB3	0.512	184	36.8	20	0.000	0.000
BBN2002C	0.499	142	28.4	18	0.182	0.087
isi02	0.498	149	29.8	15	0.385	0.109
limsiQalir2	0.497	133	26.6	11	0.188	0.196
ali2002b	0.496	181	36.2	15	0.156	0.848
ibmsqa02c	0.455	145	29.0	44	0.224	0.239
FDUT11QA1	0.434	124	24.8	6	0.139	0.957
aranea02a	0.433	152	30.4	36	0.235	0.174
nuslamp2002	0.396	105	21.0	17	0.000	0.000
pqas22	0.358	133	26.6	11	0.145	0.674

Conclusions

- Deep processing for QA
  - Exploits parsing, semantics, anaphora, reasoning
  - Computationally expensive
    - But tractable because applied only to questions and passages
- Systems trending toward greater use of:
  - Web resources: Wikipedia, answer repositories
  - Machine Learning!

Summary

- Deep Processing techniques for NLP
  - Parsing, semantic analysis, logical forms, reference, etc
  - Create richer computational models of natural language
    - Closer to language understanding
- Shallow processing techniques have dominated many areas
  - IR, QA, MT, WSD, etc
    - More computationally tractable, fewer required resources
- Deep processing techniques experience resurgence
  - Some big wins – e.g. QA
  - Improved resources: treebanks (syntactic/discourse, FrameNet, Propbank)
  - Improved learning algorithms: structured learners, neural nets
  - Increased computation: cloud resources, Grid, etc