CS 533: Natural Language Processing

Information Extraction

Karl Stratos

Rutgers University
Information Extraction (IE)

**Goal.** Extract **structured, complete** knowledge from unstructured, incomplete text

**Example input.** *The 44th president of the US takes the oath of office administered by Chief Justice at the Capitol, January 20, 2009.*

---

**What is this text about?**

**Desired information.**

- What **entities** are involved?
- What are their **relations** to each other (if any)?
- What larger **events** are taking place?
- Other domain-specific things (time, price, sentiment, etc.)
The 44th president of the US takes the oath of office administered by Chief Justice at the Capitol, January 20, 2009.
The 44th president of the US takes the oath of office administered by Chief Justice at the Capitol, January 20, 2009.

**Entities**

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(1, 5)</td>
<td>Entity: Barack Obama</td>
</tr>
<tr>
<td>2</td>
<td>(8, 10)</td>
<td>Entity: Oath of office of the President of the United States</td>
</tr>
<tr>
<td>3</td>
<td>(13, 14)</td>
<td>Entity: John Roberts</td>
</tr>
<tr>
<td>4</td>
<td>(17, 17)</td>
<td>Entity: United States Capitol</td>
</tr>
<tr>
<td></td>
<td>(2, 5)</td>
<td>Entity: President of the United States?</td>
</tr>
<tr>
<td></td>
<td>(13, 14)</td>
<td>Entity: Chief Justice of the United States?</td>
</tr>
</tbody>
</table>

**Relations**

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>takes</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>administers</td>
<td>2</td>
</tr>
</tbody>
</table>

**Location**

<table>
<thead>
<tr>
<th>Address</th>
<th>First St SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>City</td>
<td>Washington</td>
</tr>
<tr>
<td>State</td>
<td>District of Columbia</td>
</tr>
<tr>
<td>Zip Code</td>
<td>20004</td>
</tr>
</tbody>
</table>

**Time**

<table>
<thead>
<tr>
<th>Year</th>
<th>2009</th>
</tr>
</thead>
<tbody>
<tr>
<td>Month</td>
<td>January</td>
</tr>
<tr>
<td>Day</td>
<td>20</td>
</tr>
<tr>
<td>Hour</td>
<td>–</td>
</tr>
<tr>
<td>Minute</td>
<td>–</td>
</tr>
</tbody>
</table>
Information Retrieval (IR)

**Goal.** Search specific information from a set of data

- Basically document ranking (TFIDF, BM25, PageRank, ...)

IR is naturally complementary to IE

1. Retrieve relevant documents
2. Extract desired structured information from the documents
Standard IE Problems

- Named-Entity Recognition
- Entity Linking
- Coreference Resolution
- Relation Extraction
Named-Entity Recognition (NER)

- Given text, do both
  1. **Identify spans of text** that correspond to named entities
  2. **Classify** the spans into task-specific entity types (e.g., person, organization, location, etc.)

```
PER
... John Smith works at New York Times ...
```

- Recall: it can be solved as a tagging problem!

```
John/B-PER Smith/I-PER works/O at/O New/B-ORG York/I-ORG Times/I-ORG
```

- Limitation: only considers simple entity labels without considering a knowledge base (KB)
Entity Linking (EL)

- Link a span of text to an entity in KB

\[
\text{The 44th president of the US takes} \ldots \quad \rightarrow \quad \text{Entity: Barack Obama}
\]

- In simplest form: giant classification problem
  - Wikipedia: tens of millions of entities!
  - Typically approached as a pipeline: IR followed by classification

- General definition nontrivial:
  - Assume spans are given, or predict them as well?
  - Allow nested spans (president of the US, president, US, ...)?
  - Link only named entities, or also allow pronouns/verbs/others?
  - Allow nil (i.e., no entity) prediction?
Coreference Resolution (Coref)

Find all expressions that refer to the same entity in a text.

We are looking for a region of central Italy bordering the Adriatic Sea. The area is mostly mountainous and includes Mt. Corno, the highest peak of the mountain range. It also includes many sheep and an Italian entrepreneur has an idea about how to make a little...
Coref vs EL

Coref is a special case of EL if we allow linking all referring expressions (since we can cluster them based on underlying their entities)

Not strictly true under a linguistic concept called **anaphora**

- **Anaphor**: term that’s referring (“he”)
- **Antecedent**: term that’s being referred to (“Barack Obama”)

Other fine-grained linguistic concepts relevant to coref

- **Cataphora**: Anaphora in which anaphor comes before antecedent (“In his dream, Peter saw . . .”)

(Content adapted from Danchi Chen’s)
Coref Requires World Knowledge/Common Sense

Try the Winograd Schema problems: https://demo.allennlp.org/coreference-resolution/MTYwMzc0Mw==

- The city councilmen refused the demonstrators a permit because they feared violence.
- The city councilmen refused the demonstrators a permit because they advocated violence.

- The trophy didn’t fit into the suitcase because it was too large.
- The trophy didn’t fit into the suitcase because it was too small.
Relation Extraction (RE)

Extract “all relations”.

CHICAGO (AP) — 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, immediately matched the move, spokesman Tim Wagner said. United, a unit of UAL, said the increase took effect Thursday night and applies to most routes where it competes against discount carriers, such as Chicago to Dallas and Atlanta and Denver to San Francisco, Los Angeles and New York.

<table>
<thead>
<tr>
<th>Subject</th>
<th>Relation</th>
<th>Object</th>
</tr>
</thead>
<tbody>
<tr>
<td>American Airlines</td>
<td>subsidiary</td>
<td>AMR</td>
</tr>
<tr>
<td>Tim Wagner</td>
<td>employee</td>
<td>American Airlines</td>
</tr>
<tr>
<td>United Airlines</td>
<td>subsidiary</td>
<td>UAL</td>
</tr>
</tbody>
</table>

Options: regex (“X such as Y”), supervised learning (NER + pairwise classifier), and others: http://nlpprogress.com/english/relationship_extraction.html
We Need Mention Detection (MD) in All Cases!

Input
Let us talk about Obama. He has a diverse extended family and supports White Sox.

EL
Let us [talk] about [Obama] . He has a diverse [extended family] and supports [White Sox].

Coref
Let [us] talk about [Obama] . [He] has a diverse extended family and supports White Sox.

Coref + RE
Let [us] talk about [Obama] . [He] has a diverse extended family and supports [White Sox].

(2, is_fan_of, 3)
Option 1: Use an Off-the-Shelf Detector

- **NER tagger: named-entities**
  
  Let us talk about [Obama]\textsubscript{PER} . He has a diverse extended family and supports [White Sox]\textsubscript{ORG} .

- **POS tagger: pronouns, verbs**
  
  Let us \textit{talk} about Obama . \textit{He} has a diverse extended family and supports White Sox .

- **Syntactic chunker/parser: noun phrases**
  
  Let us talk about Obama . He has a diverse \textit{extended family} and supports White Sox .

Follow by training a model on top of detected (filtered) mentions
Option 2: Avoid the Problem

- Just assume mention spans are always given!

- Rationale: mention boundaries are task-specific anyway
  - We’ll only focus on the hard part (e.g., disambiguation in EL)

- Can be a realistic scenario
  - User interactively highlighting a text span in an e-book reader
Option 3: Joint Model

- Learn an EL/Coref/RE/etc. model that also performs MD

- Rationale:
  - Yes, MD is task-specific
  - But actually because of that we can do better MD if we model it jointly with the task!
  - No pipeline means no unrecoverable error propagation

- Aside: NER tagger naturally models mentions and labels jointly, but limited applicability
Examples of MD Benefiting From EL

1) MD may split a larger span into two mentions of less informative entities:

- Obama’s wife gave a speech [...] 
- Federer’s coach [...] 

2) MD may split a larger span into two mentions of incorrect entities:

- Obama Castle was built in 1601 in Japan. 
- The Kennel Club is UK’s official kennel club. 
- A bird dog is a type of gun dog or hunting dog. 
- Romeo and Juliet by Shakespeare [...] 
- Natural killer cells are a type of lymphocyte 
- Mary and Max, the 2009 movie [...] 

3) MD may choose a shorter span, referring to an incorrect entity:

- The Apple is played again in cinemas. 
- The New York Times is a popular newspaper. 

4) MD may choose a longer span, referring to an incorrect entity:

- Babies Romeo and Juliet were born hours apart. 

Table 1: Examples where MD may benefit from ED and viceversa. Each wrong MD decision (underlined) can be avoided by proper context understanding. The correct spans are shown in blue.

From Kolitsas et al. (2018)
Agenda

▶ TFIDF & BM25
▶ Entity Linking
▶ Coreference Resolution
▶ Retrieval-Based Question Answering
**Task.** Represent a document so that “similar” documents are “closer” to each other than “unsimilar” ones.
Naive Bag-of-Words Representation

Document $d$ represented as a sparse binary vector $v \in \{0, 1\}^{|V|}$

$$v_t = [[t \in d]] = \begin{cases} 1 & \text{if } t \text{ appears in the document} \\ 0 & \text{otherwise} \end{cases}$$

**Hamming distance.** Documents $v, v' \in \{0, 1\}^{|V|}$

$$\text{HammingDistance}(v, v') = \sum_{t \in V: v_t \neq v'_t} 1$$

Every term type $t \in V$ is weighted equally
Given a set of $N$ documents $D$, each document $d \in D$ is represented as a sparse vector $v \in \mathbb{R}^{|V|}$ where

$$v_t = \left[ [t \in d] \right] \times \log \frac{N}{|\{d' \in D : t \in d'\}|} = \text{tf}(t, d) \times \text{idf}_D(t)$$

1. Term frequency $\text{tf}(t, d)$: 1 if $t \in d$, 0 otherwise
   Alternatively $\text{tf}(t, d) = \text{count}(t, d)$

2. Inverse document frequency $\text{idf}_D(t)$: large if $t$ appears in few documents

**Intuition.** A term $t$ in a document $d$ is significant if $t$ appears frequently in $d$ but doesn’t appear all the time in other documents.
Similarity and Distance Under TFIDF Representations

If $v, v'$ are TFIDF representations of documents $d, d'$,

$$v^\top v' = \sum_{t \in V} \text{tf}(t, d) \times \text{tf}(t, d') \times \text{idf}_D(t)^2$$

**Cosine distance.** Documents $v, v' \in \mathbb{R}^{|V|}$

$$\text{CosineDistance}(v, v') = 1 - \cos(v, v') = 1 - \frac{v^\top v'}{||v|| \cdot ||v'||}$$

Every term type $t \in V$ in every document $d \in D$ weighted differently
**Claim.** Define term-document distribution \( p(t, d) = p(d)p(t|d) \) by \( p(d) = 1/N \) and \( p(t|d) = [[t \in d]] \). The mutual information between random term \( \tau \in V \) and document \( \delta \in D \) is

\[
I(\tau, \delta) = \frac{1}{N} \sum_{d \in D, t \in V} \text{tf}(t, d) \times \text{idf}_D(t)
\]

So the TFIDF weight for term \( t \) in document \( d \) can be viewed as how much it contributes to the general amount of information gained about a document given a term.
Proof

By the Bayes rule we have for all $t \in V$

$$p(d|t) = \frac{p(t|d)}{\sum_{d' \in D} p(t|d')} = \begin{cases} \frac{1}{|\{d' \in D : t \in d'\}|} & \text{if } t \in d \\ 0 & \text{otherwise} \end{cases} \quad \forall d \in D$$

Then for any document $d \in D$ and $t \in V$

$$\log \frac{p(d|t)}{p(d)} = \begin{cases} \text{idf}_D(t) & \text{if } t \in d \\ 0 & \text{otherwise} \end{cases}$$

Hence using $p(t|d) = \text{tf}(t, d)$ (under binary term frequency)

$$I(\tau, \delta) = \sum_{d \in D, t \in V} p(t, d) \log \frac{p(d|t)}{p(d)} = \frac{1}{N} \sum_{d \in D, t \in V} \text{tf}(t, d) \text{idf}_D(t)$$
BM25 Score

- TFIDF score with smoothing + document length modeling
- Query $q$: list of $n$ terms
- BM25 score of a document $d$ for $q$

$$BM25(d, q) = \sum_{t \in q} tf^{BM25}(t, d) \times idf^D_{BM25}(t)$$

where for some $k, b$ and average document length $L$ in $D$

$$tf^{BM25}(t, d) = \frac{\text{count}(t, d)(k + 1)}{\text{count}(t, d) + k(1 - b + b(|d| / L))}$$

$$idf^D_{BM25}(t) = \log \frac{N - |\{d' \in D : t \in d'\}| + 0.5}{|\{d' \in D : t \in d'\}| + 0.5}$$

- Currently the go-to choice for IR
Agenda

- TFIDF & BM25
- Entity Linking
- Coreference Resolution
- Retrieval-Based Question Answering
Setting

Knowledge base KB: set of entities/events of interest

Assume candidate generator $C(m) \subset KB$ that maps any contextual mention $m$ to a set of candidate entities

Assume mention boundaries are provided for simplicity

Goal: map $m$ to correct entity in $C(m)$

$m = [\text{India}]$ plays a match in England today $\rightarrow$

\begin{align*}
c_1 & = \text{India(Country)} \\
c_2 & = \text{India(Album)} \\
c_3 & = \text{India_cricket_team} \\
c_4 & = \text{La_India(Singer)}
\end{align*}
Candidate Generation

Conditional distribution over entities given mention span can be estimated from hyperlinks (e.g., in Wikipedia, web crawl)

\[ p(\text{India.cricket.team} | \text{India}) \propto \text{count}(\text{India} \rightarrow \text{India.cricket.team}) \]

Strong baseline for linking named entities (in-KB accuracy > 70% on AIDA-B test set)

Limitations

- Mostly available only for named entities
- Cannot leverage additional information like mention context or entity information in KB
Ranking Model

▶ Want to avoid doing softmax over entire KB (too large)

\[ p(\cdot|m) = softmax(\text{enc}(m)) \in [0, 1]^{|KB|} \]

▶ Instead do softmax over candidates \( c_1 \ldots c_M \in C(m) \)

\[ p(\cdot|m) = softmax(\text{score}(m, c_1) \ldots \text{score}(m, c_M)) \in [0, 1]^M \]

▶ Mention-entity score can be parameterized freely, e.g.,

\[ \text{score}(m, e) = \cos(\text{enc}(m), \text{emb}(e)) \in [-1, 1] \]

where

\[ \text{emb}(e) \in \mathbb{R}^d \quad \text{entity embedding for each } e \in \text{KB} \]

\[ \text{enc}(m) \in \mathbb{R}^d \quad \text{contextual mention encoder} \]

▶ Given annotated links, the model can be trained by maximizing log likelihood
Going Beyond Static Entity Embeddings

- Each $e \in KB$ is associated with a description $\text{desc}(e)$

  $\text{desc}(\text{India}) = \text{India is the fourth studio album by Brazilian singer Gal Costa, released on 1973 by Philips Records}$

- Making $\text{score}(m, e)$ a function of $\text{desc}(e)$ will make model read and reason with entity descriptions
  - In particular handle unseen entities at test time (as long as descriptions are provided)
- Example score function (Logeswaran et al., 2019)

  $$\text{score}(m, e) = \text{BERT}(m, \text{desc}(e))$$
Zero-Shot EL by Reading Entity Descriptions (Logeswaran et al., 2019)
Agenda

- TFIDF & BM25
- Entity Linking
- Coreference Resolution
- Retrieval-Based Question Answering
Coref

- Goal: cluster all mentions of entities
  
  "I voted for Nader because he was most aligned with my values," she said.

  Coreference Cluster 1
  Coreference Cluster 2

- Many different approaches (rule-based, mention pair, mention ranking, clustering-based)

- We’ll focus on a particular mention-ranking model (Lee et al., 2017) that jointly performs MD

- Slides in this section are made by Danchi Chen
Mention-Ranking Models

- Assign each mention its highest scoring candidate antecedent according to the model
- Add a dummy NA mention to decline linking the current mention to anything ("singleton" or "first" mention)

\[
\begin{align*}
p(NA, she) &= 0.1 \\
p(I, she) &= 0.5 \\
p(Nader, she) &= 0.1 \\
p(he, she) &= 0.1 \\
p(my, she) &= 0.2
\end{align*}
\]

Apply a softmax over the scores for candidate antecedents so probabilities sum to 1
Mention-Ranking Models

- **Training time:** only clustering information is observed (no annotation of “antecedent”), so we optimize the marginal log-likelihood of all the correct antecedents.

\[
J = \sum_{i=2}^{N} - \log \left( \sum_{j=1}^{i-1} \mathbb{1}(y_{ij} = 1) p(m_j, m_i) \right)
\]

Iterate over all the mentions in the document. Usual trick of taking negative log to go from likelihood to loss.

- **Testing time:** same as mention-pair but we only pick one antecedent for each mention
End-to-End Coreference Resolution (Lee et al., 2017)

- A mention-ranking model
- Joint mention detection and clustering — so you don’t need an additional mention detector (parser/part-of-speech tagger)

\[
J = \sum_{i=2}^{N} - \log \left( \sum_{j=1}^{i-1} \mathbb{1}(y_{ij} = 1) p(m_j, m_i) \right)
\]

Iterate over all the mentions in the document
Usual trick of taking negative log to go from likelihood to loss

We consider all the possible spans + \{NA\}

\[
p(m_j, m_i) = \frac{\exp(s(m_j, m_i))}{\sum_{j' < i} \exp(m_{j'}, m_i)}
\]

\[
N = \frac{T(T + 1)}{2} + 1
\]

T: number of words

(Lee et al, 2017): End-to-end Neural Coreference Resolution
End-to-End Coreference Resolution (Lee et al., 2017)

\[ s(i, j) = s_m(i) + s_m(j) + s_a(i, j) \]

- Are spans \( i \) and \( j \) coreferent mentions?
- Is \( i \) a mention?
- Is \( j \) a mention?
- Do they look coreferent?

\[ s_m(i) = w_m \cdot \text{FFNN}_m(g_i) \]
\[ s_a(i, j) = w_a \cdot \text{FFNN}_a([g_i, g_j, g_i \circ g_j, \phi(i, j)]) \]

Let’s compute a vector representation \( g_i \in \mathbb{R}^d \) for each span \( i \)

\( \phi(i, j) \): manual features such as speaker/gender information
End-to-End Coreference Resolution (Lee et al., 2017)

Span representation: \( g_i = [x_{\text{START}(i)}^*, x_{\text{END}(i)}^*, \hat{x}_i, \phi(i)] \)

- BILSTM hidden states for span’s start and end
- Attention-based representation (details next slide) of the words in the span
- Additional features
End-to-End Coreference Resolution (Lee et al., 2017)

Attention scores:
\[ \alpha_t = w_\alpha \cdot \text{FFNN}_\alpha(x_t^*) \]

dot product of weight vector and transformed hidden state

Attention distribution:
\[ a_{i,t} = \frac{\exp(\alpha_t)}{\sum_{k=\text{START}(i)}^{\text{END}(i)} \exp(\alpha_k)} \]

just a softmax over attention scores for the span

Final representation:
\[ \hat{x}_i = \sum_{t=\text{START}(i)}^{\text{END}(i)} a_{i,t} \cdot x_t \]

Attention-weighted sum of word embeddings

\[ g_i = [x_{\text{START}(i)}^*, x_{\text{END}(i)}^*, \hat{x}_i, \phi(i)] \]

hidden states for span’s start and end

Attention-based representation

Represents the context to the left and right of the span

Represents the span itself

Represents other information not in the text

Additional features
Computational Complexity of Exhaustive MD

- $O(T^2)$ spans in a document of length $T$
- $O(T^4)$ span pairs in a document of length $T$
  - Too expensive
  - Aggressive pruning based on model’s own score ranking

- Aside: same exhaustive MD approach has been applied to EL End-to-End Neural Entity Linking (Kolitsas et al., 2018)
  - Less computational costs (no pairs, mentions filtered by entity dictionary)
Evaluation

- You need to get both “mentions” and “clusters” correctly.
- Standard practice: we use 3 types of metrics
  - B³: mention-based
  - MUC: link-based (pair of mentions)
  - CEAF: entity-based
- .. and finally take the average of these 3 F1 scores
$B^3$ Evaluation Metric

- For each mention in the reference chain, compute a precision and a recall (e.g., # of mentions in the same reference chain with the current mention)
- The final precision/recall is an average of all the mentions
Performance

- Evaluation on English Ontonotes (CoNLL-2012 Shared Task)
- #Train: 2,802 / #Dev: 343 / #Test: 348 documents

<table>
<thead>
<tr>
<th></th>
<th>MUC</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
<td>R</td>
<td>F1</td>
<td>P</td>
<td>R</td>
<td>F1</td>
<td>P</td>
<td>R</td>
<td>F1</td>
<td>Avg. F1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lee et al. (2017) (single model)</td>
<td>78.4</td>
<td>73.4</td>
<td>75.8</td>
<td>68.6</td>
<td>61.8</td>
<td>65.0</td>
<td>62.7</td>
<td>59.0</td>
<td>60.8</td>
<td>67.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clark and Manning (2016a)</td>
<td>79.2</td>
<td>70.4</td>
<td>74.6</td>
<td>69.9</td>
<td>58.0</td>
<td>63.4</td>
<td>63.5</td>
<td>55.5</td>
<td>59.2</td>
<td>65.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clark and Manning (2016b)</td>
<td>79.9</td>
<td>69.3</td>
<td>74.2</td>
<td>71.0</td>
<td>56.5</td>
<td>63.0</td>
<td>63.8</td>
<td>54.3</td>
<td>58.7</td>
<td>65.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wiseman et al. (2016)</td>
<td>77.5</td>
<td>69.8</td>
<td>73.4</td>
<td>66.8</td>
<td>57.0</td>
<td>61.5</td>
<td>62.1</td>
<td>53.9</td>
<td>57.7</td>
<td>64.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wiseman et al. (2015)</td>
<td>76.2</td>
<td>69.3</td>
<td>72.6</td>
<td>66.2</td>
<td>55.8</td>
<td>60.5</td>
<td>59.4</td>
<td>54.9</td>
<td>57.1</td>
<td>63.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fernandes et al. (2014)</td>
<td>75.9</td>
<td>65.8</td>
<td>70.5</td>
<td>77.7</td>
<td>65.8</td>
<td>71.2</td>
<td>43.2</td>
<td>55.0</td>
<td>48.4</td>
<td>63.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clark and Manning (2015)</td>
<td>76.1</td>
<td>69.4</td>
<td>72.6</td>
<td>65.6</td>
<td>56.0</td>
<td>60.4</td>
<td>59.4</td>
<td>53.0</td>
<td>56.0</td>
<td>63.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Martschat and Strube (2015)</td>
<td>76.7</td>
<td>68.1</td>
<td>72.2</td>
<td>66.1</td>
<td>54.2</td>
<td>59.6</td>
<td>59.5</td>
<td>52.3</td>
<td>55.7</td>
<td>62.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Durrett and Klein (2014)</td>
<td>72.6</td>
<td>69.9</td>
<td>71.2</td>
<td>61.2</td>
<td>56.4</td>
<td>58.7</td>
<td>56.2</td>
<td>54.2</td>
<td>55.2</td>
<td>61.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Björkelund and Kuhn (2014)</td>
<td>74.3</td>
<td>67.5</td>
<td>70.7</td>
<td>62.7</td>
<td>55.0</td>
<td>58.6</td>
<td>59.4</td>
<td>52.3</td>
<td>55.6</td>
<td>61.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Durrett and Klein (2013)</td>
<td>72.9</td>
<td>65.9</td>
<td>69.2</td>
<td>63.6</td>
<td>52.5</td>
<td>57.5</td>
<td>54.3</td>
<td>54.4</td>
<td>54.3</td>
<td>60.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

“Eash Victories and Uphill Battles in Coreference Resolution”
Performance

Avg. F1

<table>
<thead>
<tr>
<th>Model</th>
<th>F1 Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Lee et al, 2017)</td>
<td>67.2</td>
</tr>
<tr>
<td>(Lee et al, 2018)</td>
<td>71.0</td>
</tr>
<tr>
<td>SpanBERT (Joshi &amp; Chen, 2019)</td>
<td>79.6</td>
</tr>
</tbody>
</table>
Agenda

- TFIDF & BM25
- Entity Linking
- Coreference Resolution
- Retrieval-Based Question Answering
Retrieval-Based Question Answering (QA)

▶ Goal: answer a question by consulting a KB (e.g., Wikipedia)

\[ q = \text{What does the ZIP in ZIP code stand for?} \]
\[ a = \text{Zone Improvement Plan} \]

where the answer string is a span in some text block \( b \) in KB
\[ b = \ldots \text{The term ‘ZIP’ is an acronym for Zone Improvement Plan} \ldots \]

▶ Approach: IR + IE

1. Retrieve \( K \) candidate blocks for the question \( C(q) \subset \text{KB} \) (e.g., BM25)
2. Model computes the probability of span \((i, j)\) being the answer string. Objective function at \((q, a)\)

\[
J(q, a) = \sum_{b \in C(q)} \sum_{1 \leq i \leq j \leq |b|: b_{i:j} = a} \log p(b_{i:j} | q, b)
\]
Example Model (Lee et al., 2019)

For any span $s$ in block $b$,

$$p(s|q, b) = \frac{\exp(\text{score}(q, b, s))}{\sum_{s'} \exp(\text{score}(q, b, s'))}$$

where the joint score of question $q$, block $b$, and span $s \subset b$ is computed by running BERT on $(q, b)$ and taking the start/end embeddings corresponding to $s$

$$\text{score}(q, b, s) = \text{FF} \left( \begin{bmatrix} \text{BERT}(q, b)(\text{start}(s)) \\ \text{BERT}(q, b)(\text{end}(s)) \end{bmatrix} \right)$$
Joint Retrieval + QA (Lee et al., 2019)

▸ Instead of pipeline, we can learn the model to do IR+QA jointly
▸ In addition to $p(s|q, b)$, the model additionally defines

$$p(b|q) = \frac{\exp(score(q, b))}{\sum_{b'} \exp(score(q, b'))}$$

where the joint score of question $q$ and block $b$ is computed by running BERTs on $q$ and $b$ and taking the dot product between their CLS embeddings

$$score(q, b) = BERT(q)([CLS])^\top BERT(b)([CLS])$$

▸ Can be (+ need to be) pretrained
▸ Additional objective term at $(q, a)$

$$\sum_{b \in C'(q): a \in b} \log p(b|q)$$

where candidates $C''(q) \subset KB$ retrieved under model’s own scores
Computation Graph

[Diagram showing a computation graph with BERT nodes and related text nodes.]

- BERT_Q(q)
  - [CLS] What does the zip in zip code stand for? [SEP]
  - S_retr(0, q)
  - S_retr(1, q)
  - S_retr(2, q)
  - S_retr(..., q)

- BERT_B(0)
  - [CLS] The term ‘ZIP’ is an acronym for Zone Improvement Plan... [SEP]

- BERT_B(1)
  - [CLS] group of zebras are referred to as a herd or dazzle... [SEP]

- BERT_B(2)
  - [CLS] ZIPs for other operating systems may be preceded by... [SEP]

- BERT_R(q, 0)
  - [CLS] What does the zip in zip code stand for? [SEP]
  - S_read(0, “The term”, q)
  - S_read(0, “Zone Improvement Plan”, q)
  - S_read(0, ..., q)

- BERT_R(q, 2)
  - [CLS] ZIPs for other operating systems may be preceded by... [SEP]
  - S_read(2, “ZIPs”, q)
  - S_read(2, “operating systems”, q)
  - S_read(2, ..., q)
## Performance

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Train</th>
<th>Dev</th>
<th>Test</th>
<th>Example Question</th>
<th>Example Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Natural Questions</td>
<td>79168</td>
<td>8757</td>
<td>3610</td>
<td>What does the zip in zip code stand for?</td>
<td>Zone Improvement Plan</td>
</tr>
<tr>
<td>WebQuestions</td>
<td>3417</td>
<td>361</td>
<td>2032</td>
<td>What airport is closer to downtown Houston?</td>
<td>William P. Hobby Airport</td>
</tr>
<tr>
<td>CuratedTrec</td>
<td>1353</td>
<td>133</td>
<td>694</td>
<td>What metal has the highest melting point?</td>
<td>Tungsten</td>
</tr>
<tr>
<td>TriviaQA</td>
<td>78785</td>
<td>8837</td>
<td>11313</td>
<td>What did L. Fran Baum, author of The Wonderful Wizard of Oz, call his home in Hollywood?</td>
<td>Ozcot</td>
</tr>
<tr>
<td>SQuAD</td>
<td>78713</td>
<td>8886</td>
<td>10570</td>
<td>Other than the Automobile Club of Southern California, what other AAA Auto Club chose to simplify the divide?</td>
<td>California State Automobile Association</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Train</th>
<th>Dev</th>
<th>Test</th>
<th>Model</th>
<th>BM25 +BERT</th>
<th>NNLMT +BERT</th>
<th>ELMO +BERT</th>
<th>ORQA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Natural Questions</td>
<td></td>
<td></td>
<td></td>
<td>Dev</td>
<td>24.8</td>
<td>3.2</td>
<td>3.6</td>
<td>31.3</td>
</tr>
<tr>
<td>WebQuestions</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>20.8</td>
<td>9.1</td>
<td>17.7</td>
<td>38.5</td>
</tr>
<tr>
<td>CuratedTrec</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>27.1</td>
<td>6.0</td>
<td>8.3</td>
<td>36.8</td>
</tr>
<tr>
<td>TriviaQA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>47.2</td>
<td>7.3</td>
<td>6.0</td>
<td>45.1</td>
</tr>
<tr>
<td>SQuAD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>28.1</td>
<td>2.8</td>
<td>1.9</td>
<td>26.5</td>
</tr>
<tr>
<td></td>
<td>Test</td>
<td></td>
<td></td>
<td>Natural Questions</td>
<td>26.5</td>
<td>4.0</td>
<td>4.7</td>
<td>33.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>WebQuestions</td>
<td>17.7</td>
<td>7.3</td>
<td>15.6</td>
<td>36.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>CuratedTrec</td>
<td>21.3</td>
<td>4.5</td>
<td>6.8</td>
<td>30.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>TriviaQA</td>
<td>47.1</td>
<td>7.1</td>
<td>5.7</td>
<td>45.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>SQuAD</td>
<td>33.2</td>
<td>3.2</td>
<td>2.3</td>
<td>20.2</td>
</tr>
</tbody>
</table>