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.)

Example Output

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

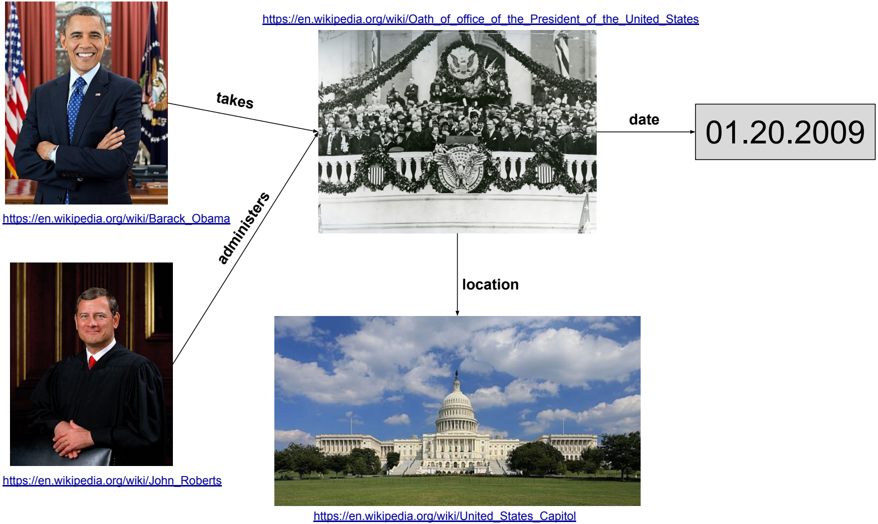


Table Form

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

Entities

1	(1, 5)	Entity:Barack_Obama
2	(8, 10)	Entity:Oath_of_office_of_the_President_of_the_United_States
3	(13, 14)	Entity:John_Roberts
4	(17, 17)	Entity:United_States_Capitol
	(2, 5)	Entity:President_of_the_United_States?
	(13, 14)	Entity:Chief_Justice_of_the_United_States?

Relations

1	takes	2
3	administers	2

Location

Address	First St SE
City	Washington
State	District of Columbia
Zip Code	20004

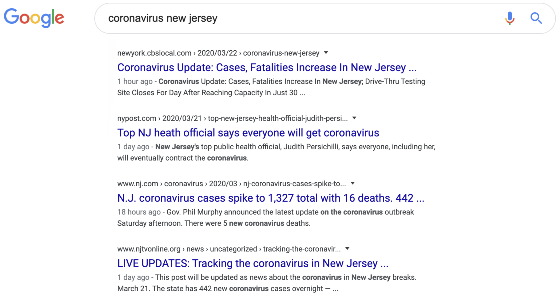
Time

Year	2009
Month	January
Day	20
Hour	–
Minute	–

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.)

$$\ldots\ \overbrace{\texttt{John Smith}}^{\text{PER}}\ \texttt{works at}\ \overbrace{\texttt{New York Times}}^{\text{ORG}}\ \ldots$$

- 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)

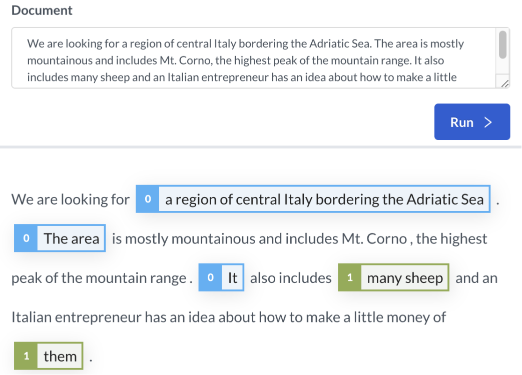
- Link a **span of text** to an **entity in KB**

$$\underbrace{\textit{The } {\color{red}\textit{44th president of the US}} \textit{ takes} \ldots}_{\text{contextualized mention}} \quad \longrightarrow \quad \underbrace{\text{Entity:Barack\_Obama}}_{\text{entity}}$$

- In simplest form: **giant** classification problem
  - Wikipedia: tens of millions of entities!
  - Typically approached as a pipeline: IR followed by classification
- General definition nontrivial: https://www.aclweb.org/anthology/Q15-1023.pdf
  - 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.

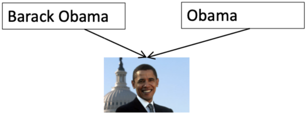


Coref vs EL

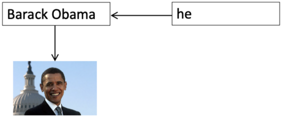
Coref is a special case of EL if we allow linking all referring expressions (since we can cluster them based on underlying their entities)

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

- **Anaphor**: term that's referring ("he")
- **Antecedent**: term that's being referred to ("Barack Obama")



vs



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 <u>feared</u> violence.
- The city councilmen refused the demonstrators a permit because they <u>advocated</u> violence.

- The trophy didn't fit into the suitcase because it was too <u>large</u>.
- The trophy didn't fit into the suitcase because it was too <u>small</u>.

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.

Subject	Relation	Object
American Airlines	subsidiary	AMR
Tim Wagner	employee	American Airlines
United Airlines	subsidiary	UAL

example from Jim Martin

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]$_{\textit{Entity:Conversation}}$ about [Obama]$_{\textit{Entity:Barack\_Obama}}$. He has a diverse [extended family]$_{\textit{Entity:Family\_of\_Barack\_Obama}}$ and supports [White Sox]$_{\textit{Entity:Chicago\_White\_Sox}}$.

Coref

Let [us]$_1$ talk about [Obama]$_2$. [He]$_2$ has a diverse extended family and supports White Sox .

Coref + RE

Let [us]$_1$ talk about [Obama]$_2$. [He]$_2$ has a diverse extended family and supports [White Sox]$_3$.

$$(2, \text{is\_fan\_of}, 3)$$

Option 1: Use an Off-the-Shelf Detector

- NER tagger: named-entities

  Let us talk about [Obama]$_{PER}$. He has a diverse extended family and supports [White Sox]$_{ORG}$.

- POS tagger: pronouns, verbs

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

- Syntactic chunker/parser: noun phrases

  Let us talk about Obama . He has a diverse [extended family]$_{NP}$ 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:
B. 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.

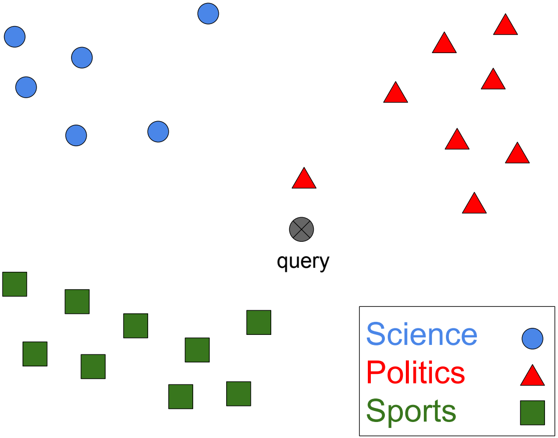
From Kolitsas et al. (2018)

Agenda

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

Document Representations

Task. Represent a document so that "similar" documents are "closer" to each other than "unsimilar" ones



Naive Bag-of-Words Representation

$$\rightarrow \quad \underbrace{(0,0,0,{\color{red}1},\ldots,0,{\color{red}1},0,\ldots,0,0)}_{\text{Vocabulary size}}$$

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

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

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

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

Every term type $t \in V$ is weighted equally

TFIDF Representation

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

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

1. Term frequency $\mathrm{tf}(t,d)$: 1 if $t \in d$, 0 otherwise

   Alternatively $\mathrm{tf}(t,d) = \textbf{count}(t,d)$
2. Inverse document frequency $\mathrm{idf}_D(t)$: large if t appears in few documents

Intution. 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',

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

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

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

Every term type $t \in V$ in every document $d \in D$ weighted differently

Connection to Mutual Information

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

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

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 $t \in V$

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

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

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

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

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

BM25 Score

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

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

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

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

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

- 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** $C(m) \subset \mathrm{KB}$ 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 $C(m)$

$c_1 =$ India(Country)

$c_2 =$ Índia(Album)

$m =$ `[India] plays a match in England today` $\longrightarrow$ $c_4 =$ La_India(Singer)

$c_3 =$ India_cricket_team

Candidate Generation

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

$$p(\text{India\_cricket\_team} \,|\, \texttt{India}) \propto \textbf{count}(\texttt{India} \mapsto \text{India\_cricket\_team})$$

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)

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

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

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

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

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

where

$$\textbf{emb}(e) \in \mathbb{R}^d \qquad \text{entity embedding for each } e \in \text{KB}$$

$$\textbf{enc}(m) \in \mathbb{R}^d \qquad \text{contextual mention encoder}$$

- Given annotated links, the model can be trained by maximizing log likelihood

Going Beyond Static Entity Embeddings

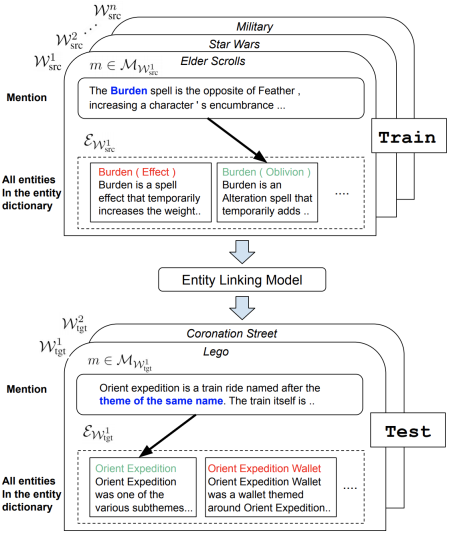
- Each $e \in \mathrm{KB}$ is associated with a description $\mathrm{desc}(e)$

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

- Making **score**(m, e) a function of $\mathrm{desc}(e)$ 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)

$$\textbf{score}(m, e) = \textbf{BERT}(m, \mathrm{desc}(e))$$

Zero-Shot EL by Reading Entity Descriptions (Logeswaran et al., 2019)

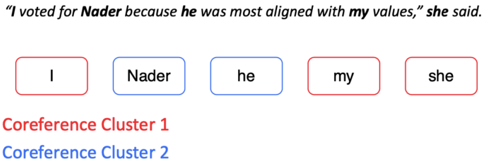


Agenda

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

Coref

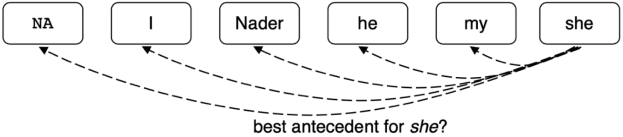
- Goal: cluster all mentions of entities



- 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)



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

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)

$$s(i,j) = s_{\mathrm{m}}(i) + s_{\mathrm{m}}(j) + s_{\mathrm{a}}(i,j)$$

Are spans *i* and *j* coreferent mentions?

Is *i* a mention?

Is *j* a mention?

Do they look coreferent?

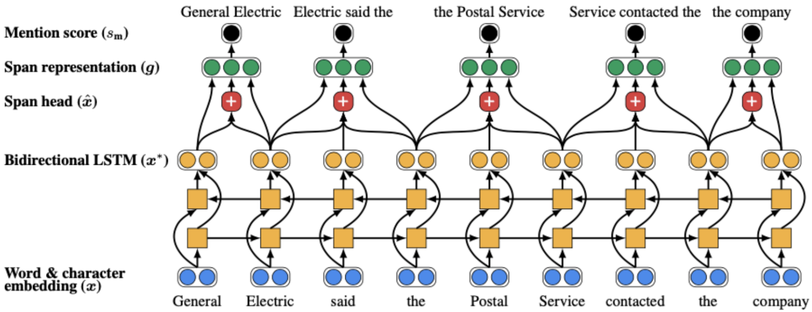
$$s_{\mathrm{m}}(i) = \boldsymbol{w}_{\mathrm{m}} \cdot \mathrm{FFNN}_{\mathrm{m}}(\boldsymbol{g}_i)$$

$$s_{\mathrm{a}}(i,j) = \boldsymbol{w}_{\mathrm{a}} \cdot \mathrm{FFNN}_{\mathrm{a}}([\boldsymbol{g}_i, \boldsymbol{g}_j, \boldsymbol{g}_i \circ \boldsymbol{g}_j, \phi(i,j)])$$

Let's compute a vector representation $\mathbf{g}_i \in \mathbb{R}^d$ for each span i

$\phi(i,j)$: manual features such speaker/gender information

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



Span representation: $\boldsymbol{g}_i = [\boldsymbol{x}^*_{\text{START}(i)}, \boldsymbol{x}^*_{\text{END}(i)}, \hat{\boldsymbol{x}}_i, \phi(i)]$

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

$$\alpha_t = \boldsymbol{w}_\alpha \cdot \mathrm{FFNN}_\alpha(\boldsymbol{x}_t^*)$$

dot product of weight vector and transformed hidden state

Attention distribution

$$a_{i,t} = \frac{\exp(\alpha_t)}{\sum_{k=\mathrm{START}(i)}^{\mathrm{END}(i)} \exp(\alpha_k)}$$

just a softmax over attention scores for the span

Final representation

$$\hat{\boldsymbol{x}}_i = \sum_{t=\mathrm{START}(i)}^{\mathrm{END}(i)} a_{i,t} \cdot \boldsymbol{x}_t$$

Attention-weighted sum of word embeddings

$$\boldsymbol{g}_i = [\boldsymbol{x}^*_{\mathrm{START}(i)}, \boldsymbol{x}^*_{\mathrm{END}(i)}, \hat{\boldsymbol{x}}_i, \phi(i)]$$

hidden states for span's start and end

Attention-based representation

Additional features

Represents the context to the left and right of the span

Represents the span itself

Represents other information not in the text

Computational Complexity of Exhaustive MD

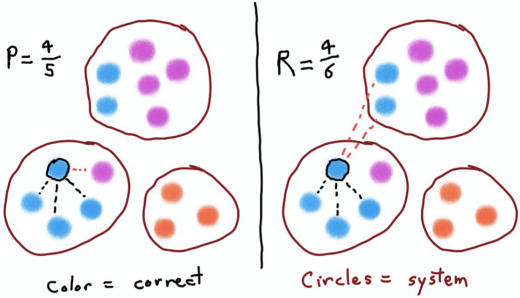
- $O(T^2)$ spans in a document of length T
- $O(T^4)$ span pairs in a document of length T
  - Too expensive
  - Aggressive pruning based on model's own score ranking
- Aside: same exhaustive MD approach has been applied to EL
  End-to-End Neural Entity Linking (Kolitsas et al., 2018)
  - Less computational 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
  - B^3: 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



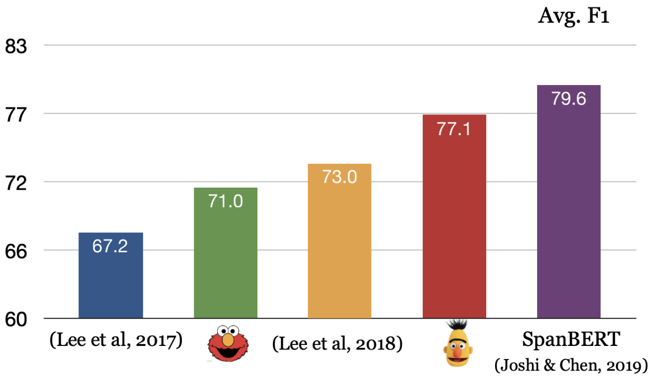
Performance

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

	MUC			B^3			$CEAF_{\phi_4}$			
	P	R	F1	P	R	F1	P	R	F1	Avg. F1
Lee et al. (2017) (single model)	78.4	73.4	75.8	68.6	61.8	65.0	62.7	59.0	60.8	67.2
Clark and Manning (2016a)	79.2	70.4	74.6	69.9	58.0	63.4	63.5	55.5	59.2	65.7
Clark and Manning (2016b)	79.9	69.3	74.2	71.0	56.5	63.0	63.8	54.3	58.7	65.3
Wiseman et al. (2016)	77.5	69.8	73.4	66.8	57.0	61.5	62.1	53.9	57.7	64.2
Wiseman et al. (2015)	76.2	69.3	72.6	66.2	55.8	60.5	59.4	54.9	57.1	63.4
Fernandes et al. (2014)	75.9	65.8	70.5	77.7	65.8	71.2	43.2	55.0	48.4	63.4
Clark and Manning (2015)	76.1	69.4	72.6	65.6	56.0	60.4	59.4	53.0	56.0	63.0
Martschat and Strube (2015)	76.7	68.1	72.2	66.1	54.2	59.6	59.5	52.3	55.7	62.5
Durrett and Klein (2014)	72.6	69.9	71.2	61.2	56.4	58.7	56.2	54.2	55.2	61.7
Björkelund and Kuhn (2014)	74.3	67.5	70.7	62.7	55.0	58.6	59.4	52.3	55.6	61.6
Durrett and Klein (2013)	72.9	65.9	69.2	63.6	52.5	57.5	54.3	54.4	54.3	60.3

"**Eash Victories** and **Uphill Battles** in Coreference Resolution"

Performance



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)

$q =$ `What does the ZIP in ZIP code stand for?`

$a =$ `Zone Improvement Plan`

where the answer string is a span in some text block b in KB

b = `... The term 'ZIP' is an acronym for Zone Improvement Plan ...`

- Approach: IR + IE
  1. Retrieve K candidate blocks for the question $C(q) \subset \mathrm{KB}$ (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,

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

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

$$\textsf{\textbf{score}}(q,b,s) = \mathrm{FF}\left(\begin{bmatrix} \textsf{\textbf{BERT}}(q,b)(\mathrm{start}(s)) \\ \textsf{\textbf{BERT}}(q,b)(\mathrm{end}(s)) \end{bmatrix}\right)$$

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

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

  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

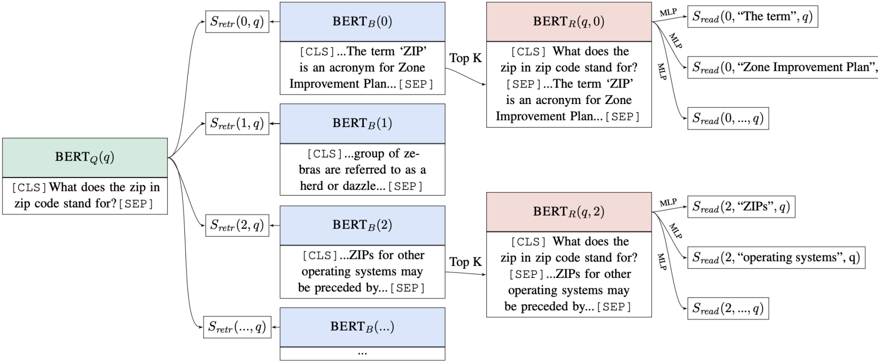
$$\textbf{score}(q,b) = \textbf{BERT}(q)([\text{CLS}])^\top \textbf{BERT}(b)([\text{CLS}])$$

- 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 $C'(q) \subset \text{KB}$ retrieved under model's own scores

Computation Graph



Performance

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

	Model	BM25 +BERT	NNLM +BERT	ELMo +BERT	ORQA
Dev	Natural Questions	24.8	3.2	3.6	**31.3**
	WebQuestions	20.8	9.1	17.7	**38.5**
	CuratedTrec	27.1	6.0	8.3	**36.8**
	TriviaQA	**47.2**	7.3	6.0	45.1
	SQuAD	**28.1**	2.8	1.9	26.5
Test	Natural Questions	26.5	4.0	4.7	**33.3**
	WebQuestions	17.7	7.3	15.6	**36.4**
	CuratedTrec	21.3	4.5	6.8	**30.1**
	TriviaQA	**47.1**	7.1	5.7	45.0
	SQuAD	**33.2**	3.2	2.3	20.2