NLP Application II

Named Entity Recognition and Entity Linking

Slides source (ANLP, David Bamman, UC Berkeley)
Slides source (Dan Roth, Ming Wei Chang and Taylor Cassidy, UPenn)

Outline

- Named entity recognition
- Entity linking

Named entity recognition

[tim cook]$_{\text{PER}}$ is the ceo of [apple]$_{\text{ORG}}$

- Identifying spans of text that correspond to typed entities

Named entity recognition

Type	Tag	Sample Categories	Example sentences
People	PER	people, characters	**Turing** is a giant of computer science.
Organization	ORG	companies, sports teams	The **IPCC** warned about the cyclone.
Location	LOC	regions, mountains, seas	The **Mt. Sanitas** loop is in **Sunshine Canyon**.
Geo-Political Entity	GPE	countries, states, provinces	**Palo Alto** is raising the fees for parking.
Facility	FAC	bridges, buildings, airports	Consider the **Golden Gate Bridge**.
Vehicles	VEH	planes, trains, automobiles	It was a classic **Ford Falcon**.

Figure 17.1 A list of generic named entity types with the kinds of entities they refer to.

ACE NER categories (+weapon)

Named entity recognition

- GENIA corpus of MEDLINE abstracts (biomedical)

We have shown that [interleukin-1]$_{\text{PROTEIN}}$ ([IL-1]$_{\text{PROTEIN}}$) and [IL-2]$_{\text{PROTEIN}}$ control [IL-2 receptor alpha (IL-2R alpha) gene]$_{\text{DNA}}$ transcription in [CD4-CD8- murine T lymphocyte precursors]$_{\text{CELL LINE}}$

protein
cell line
cell type
DNA
RNA

http://www.aclweb.org/anthology/W04-1213

BIO notation

B-PERS I-PERS O O O O B-ORG

tim cook is the ceo of apple

- **B**eginning of entity
- **I**nside entity
- **O**utside entity

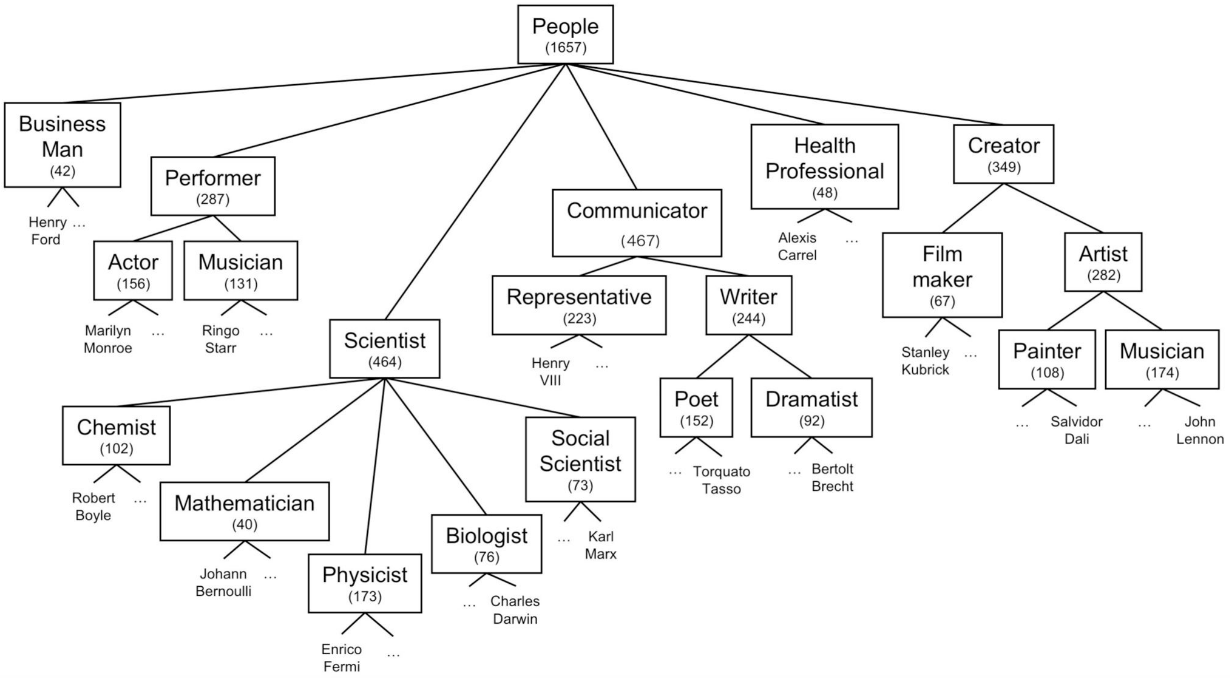
[tim cook]$_{\text{PER}}$ is the ceo of [apple]$_{\text{ORG}}$

Named entity recognition

B-PERS B-PERS

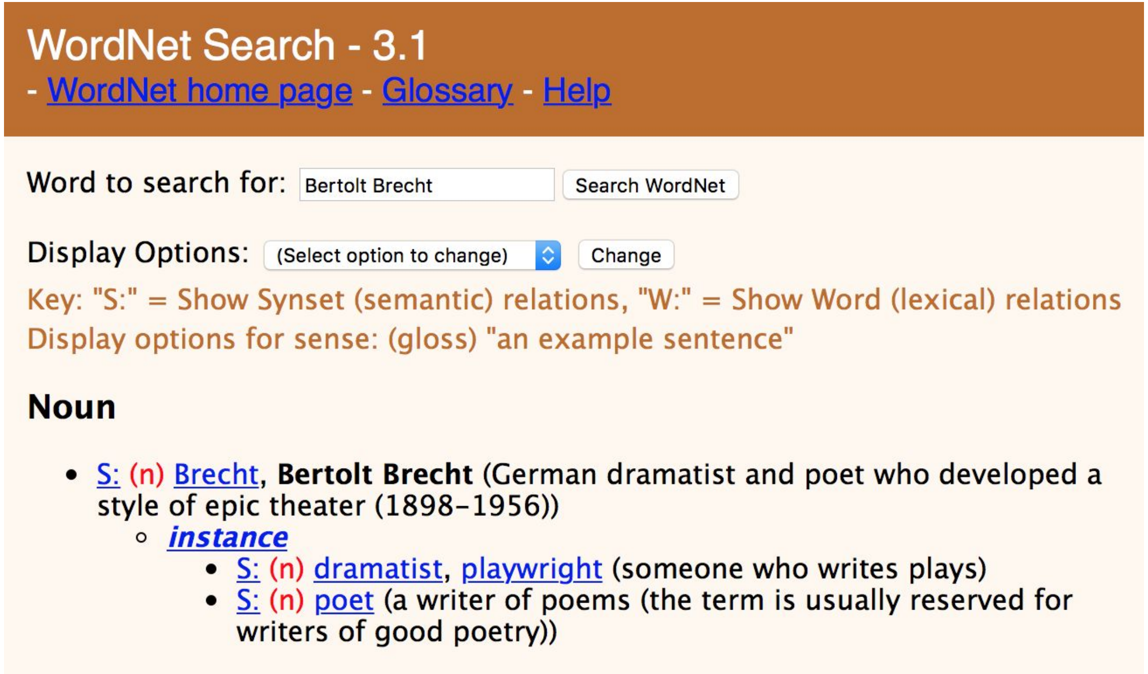
After he saw Harry Tom went to the store

Fine-grained NER



Giuliano and Gliozzo (2008)

Fine-grained NER



WordNet Search - 3.1

- WordNet home page - Glossary - Help

Word to search for: Bertolt Brecht Search WordNet

Display Options: (Select option to change) Change

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations

Display options for sense: (gloss) "an example sentence"

Noun

- S: (n) Brecht, **Bertolt Brecht** (German dramatist and poet who developed a style of epic theater (1898–1956))
  - ***instance***
    - S: (n) dramatist, playwright (someone who writes plays)
    - S: (n) poet (a writer of poems (the term is usually reserved for writers of good poetry))

Entity recognition

Person	… named after [the daughter of a Mattel co-founder] …
Organization	[The Russian navy] said the submarine was equipped with 24 missiles
Location	Fresh snow across [the upper Midwest] on Monday, closing schools
GPE	The [Russian] navy said the submarine was equipped with 24 missiles
Facility	Fresh snow across the upper Midwest on Monday, closing [schools]
Vehicle	The Russian navy said [the submarine] was equipped with 24 missiles
Weapon	The Russian navy said the submarine was equipped with [24 missiles]

ACE entity categories
https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-entities-guidelines-v6.6.pdf

Named entity recognition

- Most named entity recognition datasets have flat structure (i.e., non-hierarchical labels).

✔ [The University of California]$_{\text{ORG}}$

✖ [The University of [California]$_{\text{GPE}}$]$_{\text{ORG}}$

- Mostly fine for named entities, but more problematic for general entities:

[[John]$_{\text{PER}}$'s mother]$_{\text{PER}}$ said …

Nested NER

named	after	the	daughter	of	a	Mattel	co-founder
						B-ORG	
					B-PER	I-PER	I-PER
		B-PER	I-PER	I-PER	I-PER	I-PER	I-PER

Nested NER

"in the US Federal District Court of New Mexico."

in	O
the	B-ORG
US	I-ORG|U-GPE
Federal	I-ORG
District	I-ORG|U-GPE
Court	I-ORG
of	I-ORG
New	I-ORG|B-GPE
Mexico	L-ORG|L-GPE
.	O

B-: beginning
I-: inside)
U-: unit-length entity
L-: last
O : outside

Strakova et al. (2019)

Sequence labeling

$$x = \{x_1, \ldots, x_n\}$$

$$y = \{y_1, \ldots, y_n\}$$

- For a set of inputs x with n sequential time steps, one corresponding label y_i for each x_i
- Model correlations in the labels y.

Sequence labeling

- Feature-based models (MEMM, CRF)

identity of w_i, identity of neighboring words
embeddings for w_i, embeddings for neighboring words
part of speech of w_i, part of speech of neighboring words
base-phrase syntactic chunk label of w_i and neighboring words
presence of w_i in a **gazetteer**
w_i contains a particular prefix (from all prefixes of length ≤ 4)
w_i contains a particular suffix (from all suffixes of length ≤ 4)
w_i is all upper case
word shape of w_i, word shape of neighboring words
short word shape of w_i, short word shape of neighboring words
presence of hyphen

Figure 17.5 Typical features for a feature-based NER system.

Gazetteers

- List of place names; more generally, list of names of some typed category
- GeoNames (GEO), US SSN (PER), Getty Thesaurus of Geographic Placenames, Getty Thesaurus of Art and Architecture

Bun Cranncha
Dromore West
Dromore
Youghal Harbour
Youghal Bay
Youghal
Eochaill
Yellow River
Yellow Furze
Woodville
Wood View
Woodtown House
Woodstown
Woodstock House
Woodsgift House
Woodrooff House
Woodpark
Woodmount
Wood Lodge
Woodlawn Station
Woodlawn
Woodlands Station
Woodhouse
Wood Hill
Woodfort
Woodford River
Woodford
Woodfield House
Woodenbridge Junction Station
Woodenbridge
Woodbrook House
Woodbrook
Woodbine Hill
Wingfield House
Windy Harbour
Windy Gap
Windgap
Windfield House
Wilton House
Wilton Castle
Wilmount House
Wilmount
Wills Grove

Conditional Random Fields (CRF)

- Compute directly the posterior (p(Y|X)) of a tag sequence given the input text

$$p(Y|X) = \frac{\exp\left(\sum_{k=1}^{K} w_k F_k(X,Y)\right)}{\sum_{Y' \in \mathscr{Y}} \exp\left(\sum_{k=1}^{K} w_k F_k(X,Y')\right)}$$

- Giant version of a multinomial logistic regression
- F_k maps entire input sequence X and entire output sequence Y to a feature vector of K features (**global features**)
- W_k is the weight for each feature F_k, which are computed as a sum of local features for each position.

$$F_k(X,Y) = \sum_{i=1}^{n} f_k(y_{i-1}, y_i, X, i)$$

Linear chain CRF relies on the current and previous token predictions

In a CRF, we use features from the entire sequence (by summing the individual features at each time step)

	will $\phi(x, 1, y_1, y_0)$	to $\phi(x, 2, y_2, y_1)$	fight $\phi(x, 3, y_3, y_2)$	$\Phi(x, \text{NN TO VB})$
x_i=will ^ y_i = NN	1	0	0	1
y_{i-1}=START ^ y_i = NN	1	0	0	1
x_i=will ^ y_i = MD	0	0	0	0
y_{i-1}=START ^ y_i = MD	0	0	0	0
…				
x_i=to ^ y_i = TO	0	1	0	1
y_{i-1}=NN ^ y_i = TO	0	1	0	1
y_{i-1}=MD ^ y_i = TO	0	0	0	0
…				
x_i=fight ^ y_i = VB	0	0	1	1
y_{i-1}=TO ^ y_i = VB	0	0	1	1

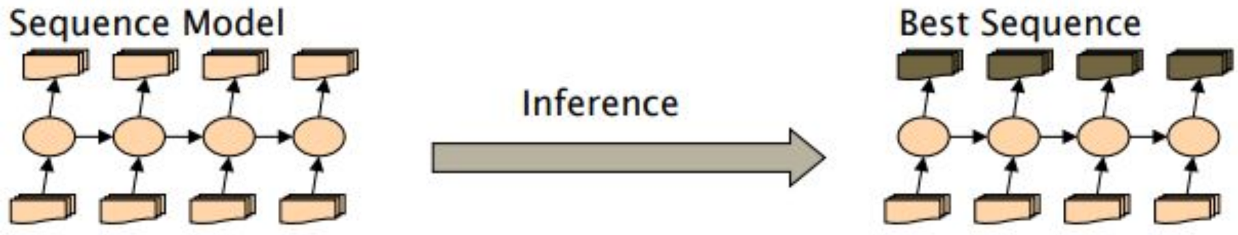
This lets us isolate the global sequence features that separate good sequences (in our training data) from bad sequences (not in our training data)

	Φ(x, NN TO VB) GOOD	Φ(x, MD TO VB) BAD
x_i=will ^ y_i = NN	1	0
y_{i-1}=START ^ y_i = NN	1	0
x_i=will ^ y_i = MD	0	1
y_{i-1}=START ^ y_i = MD	0	1
…		
y_{i-1}=NN ^ y_i = TO	1	0
y_{i-1}=MD ^ y_i = TO	0	1
x_i=to ^ y_i = TO	1	1
x_i=fight ^ y_i = VB	1	1
y_{i-1}=TO ^ y_i = VB	1	1

these are the different (and so are potentially predictive of a good label sequence)

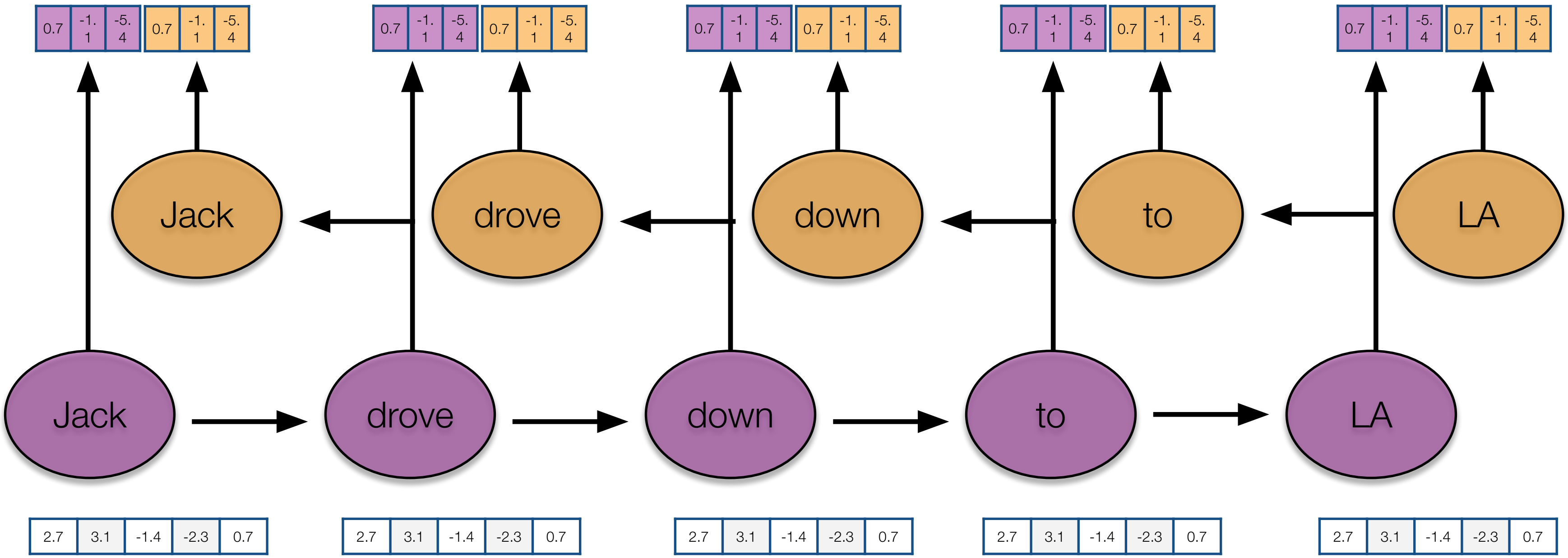
these are the same (and so are not)

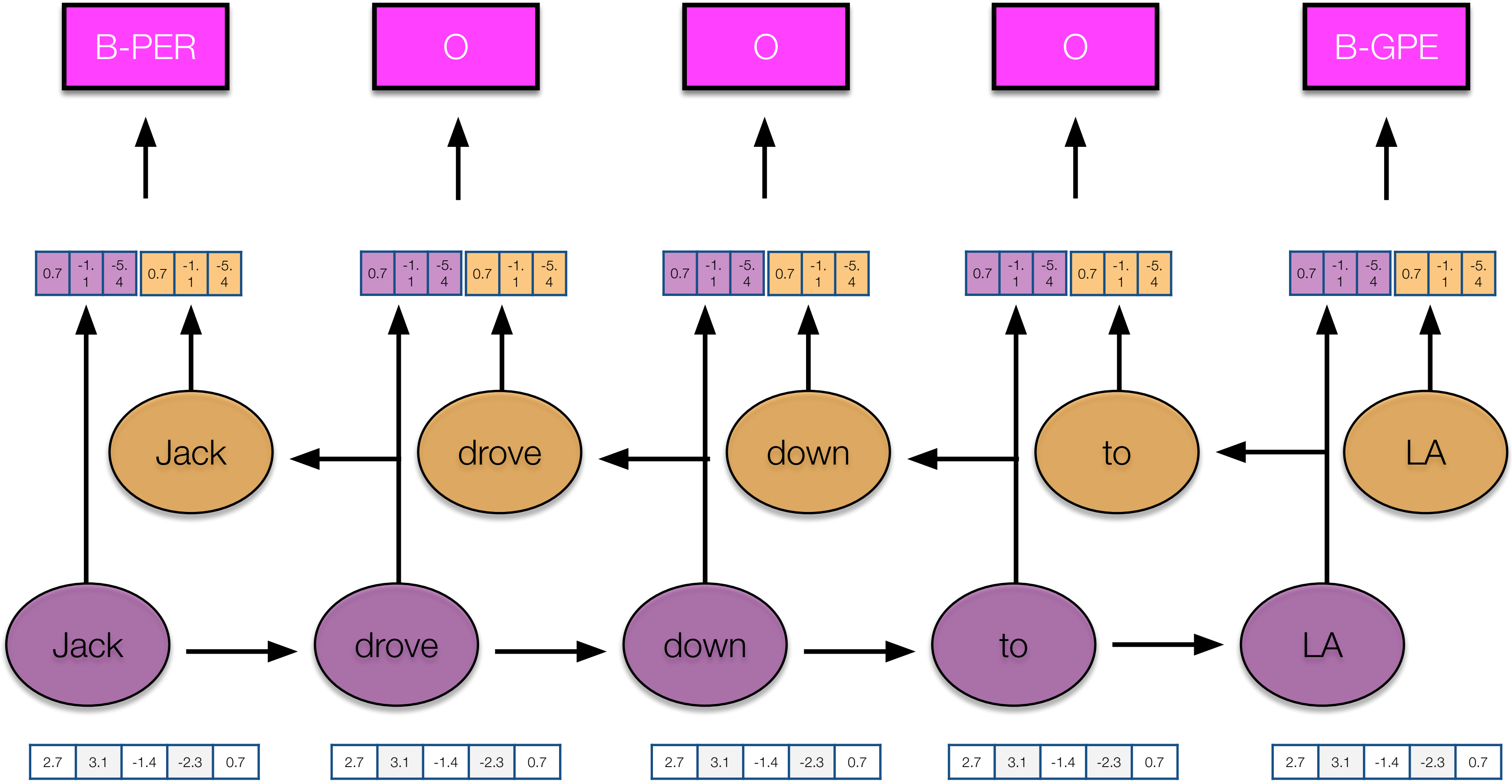
Inferences with CRFs



- Greedy inference:
  - Use our classifier at each position to assign a label (can use previous predictions).
  - Fast, no extra memory, but error cannot recover.
- Beam search:
  - At each position keep the top k complete sequences.
  - Fast, beam 3-5 similar to exact inference.
- Viterbi inference:
  - Dynamic programming, harder to implement.
  - Exact: global best sequence is returner

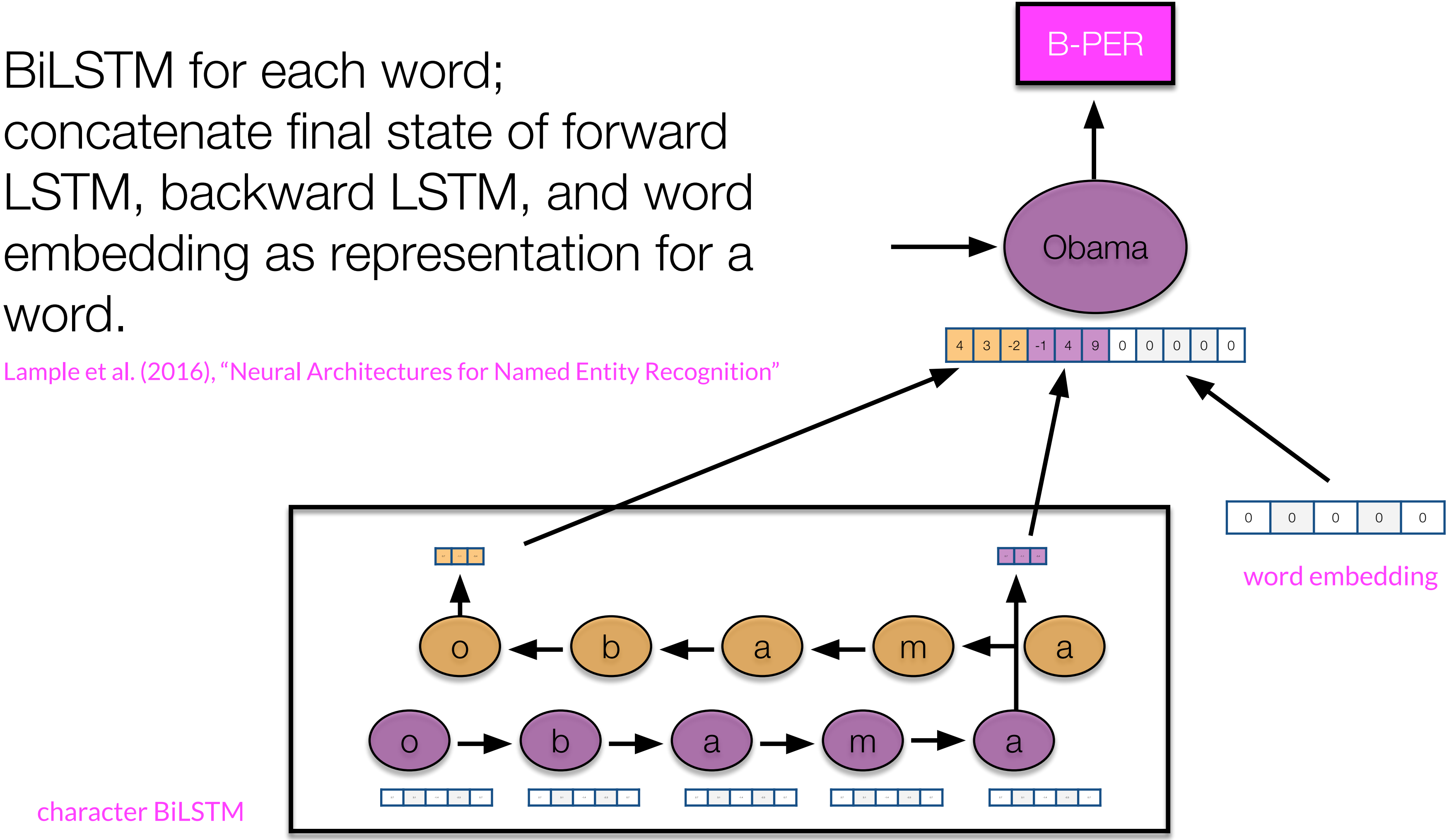
Bidirectional RNN




B-PER
O
O
O
B-GPE
0.7
-1.
1
-5.
4
Jack
drove
down
to
LA
Jack
drove
down
to
LA
2.7
3.1
-1.4
-2.3
0.7

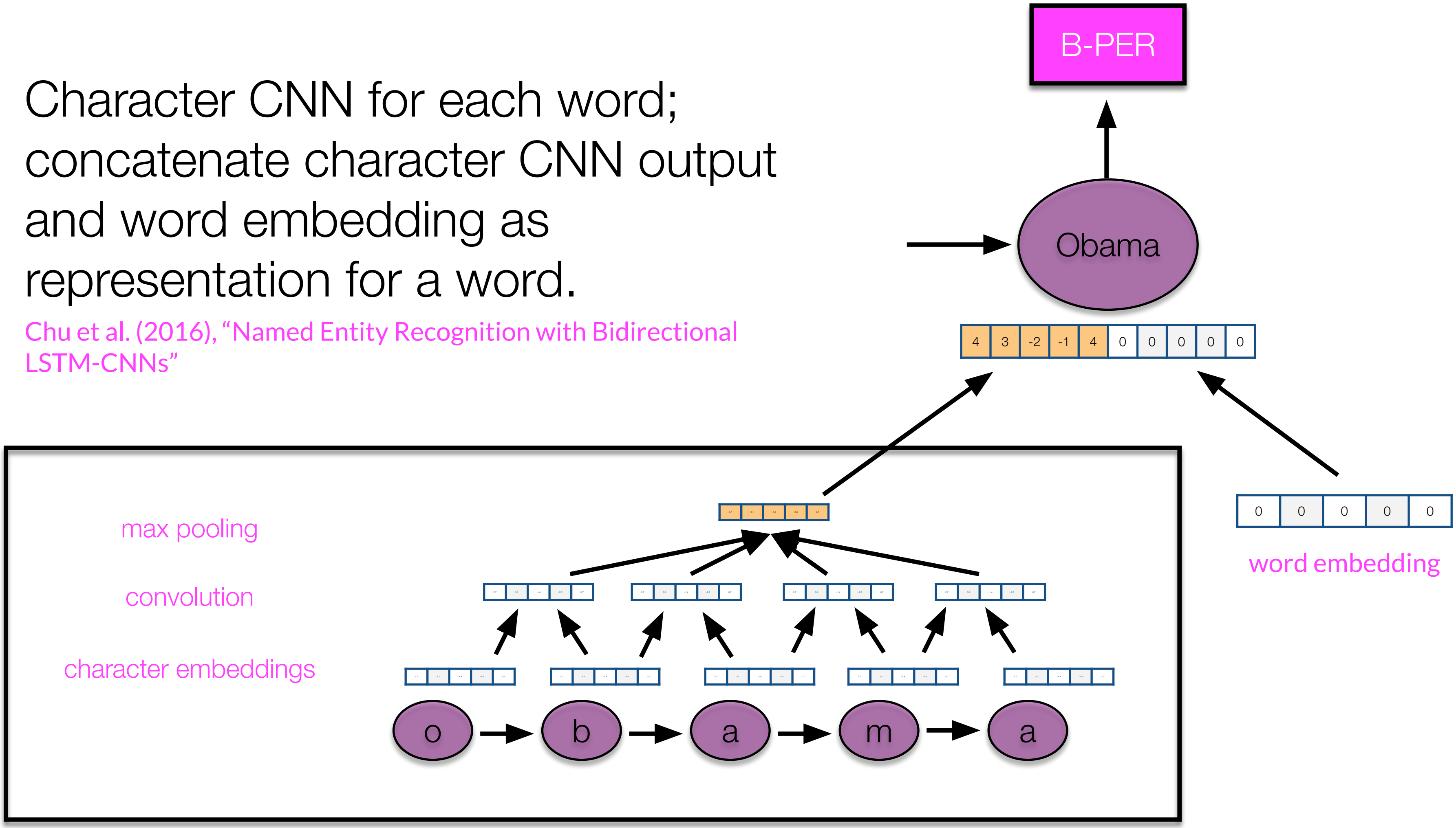
BiLSTM for each word; concatenate final state of forward LSTM, backward LSTM, and word embedding as representation for a word.

Lample et al. (2016), "Neural Architectures for Named Entity Recognition"

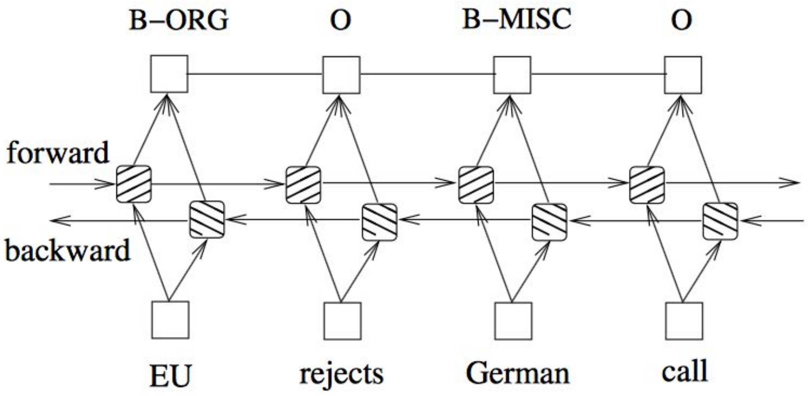


Character CNN for each word; concatenate character CNN output and word embedding as representation for a word.

Chu et al. (2016), "Named Entity Recognition with Bidirectional LSTM-CNNs"

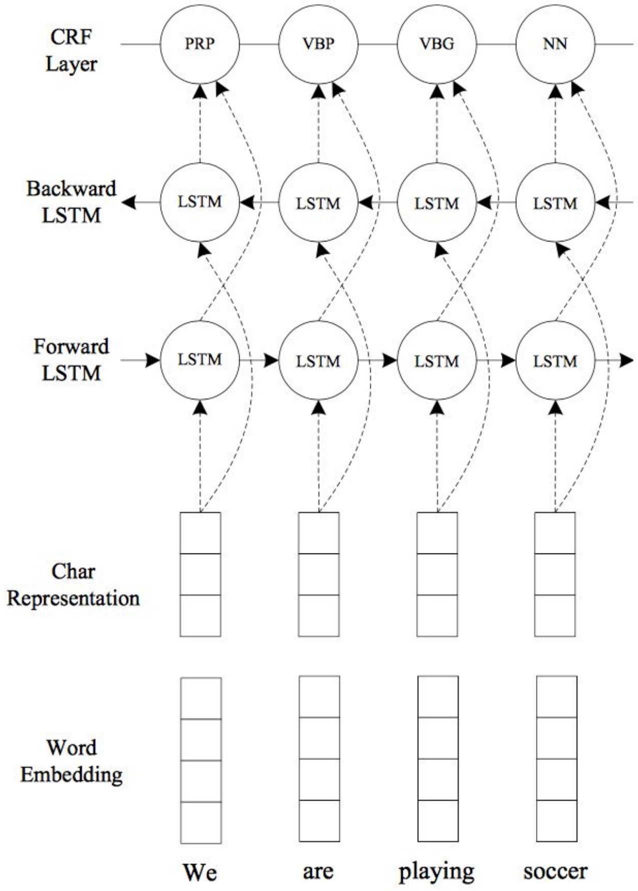


LSTM-CRF



Huang et al. 2015, "Bidirectional LSTM-CRF Models for Sequence Tagging"

Ma and Hovy (2016), “End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF”

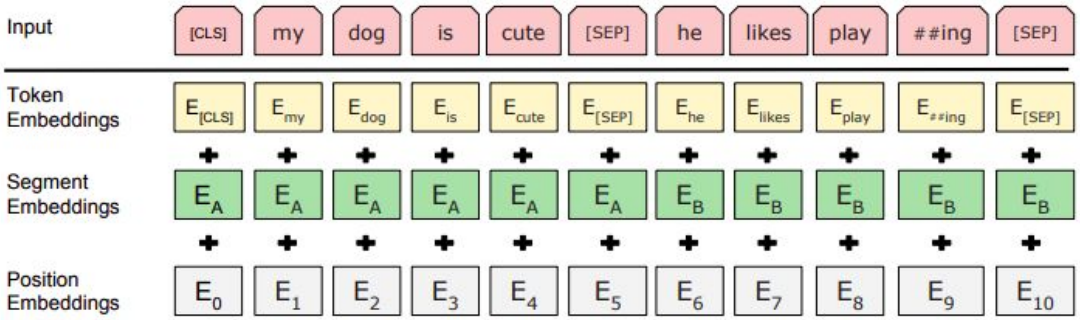


Model	POS		NER					
	Dev	Test	Dev			Test		
	Acc.	Acc.	Prec.	Recall	F1	Prec.	Recall	F1
BRNN	96.56	96.76	92.04	89.13	90.56	87.05	83.88	85.44
BLSTM	96.88	96.93	92.31	90.85	91.57	87.77	86.23	87.00
BLSTM-CNN	97.34	97.33	92.52	93.64	93.07	88.53	90.21	89.36
BRNN-CNN-CRF	97.46	97.55	94.85	94.63	94.74	91.35	91.06	91.21

Ma and Hovy (2016), “End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF”

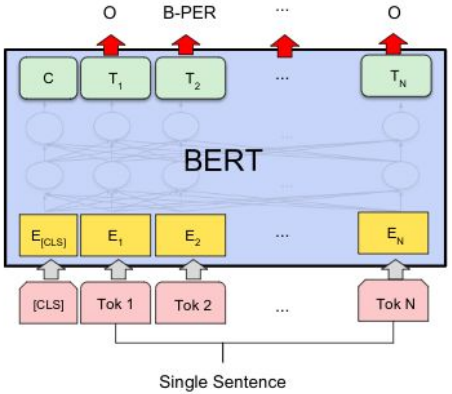
Transformers

Devlin et al. (2019), “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”



- Input representation has multiple embedding types

- Fine-tuning on single sentence tagging.
- Prediction based on first hidden layer.
- Feature based approach very competitive!



System	Dev F1	Test F1
ELMo (Peters et al., 2018a)	95.7	92.2
CVT (Clark et al., 2018)	-	92.6
CSE (Akbik et al., 2018)	-	**93.1**
Fine-tuning approach		
$\text{BERT}_{\text{LARGE}}$	96.6	92.8
$\text{BERT}_{\text{BASE}}$	96.4	92.4
Feature-based approach ($\text{BERT}_{\text{BASE}}$)		
Embeddings	91.0	-
Second-to-Last Hidden	95.6	-
Last Hidden	94.9	-
Weighted Sum Last Four Hidden	95.9	-
Concat Last Four Hidden	96.1	-
Weighted Sum All 12 Layers	95.5	-

Evaluation

	1	2	3	4	5	6	7
	tim	cook	is	the	CEO	of	Apple
gold	B-PER	I-PER	O	O	O	O	B-ORG
system	B-PER	O	O	O	B-PER	O	B-ORG

<start, end, type>

Precision	1/3
Recall	1/2

gold

<1,2,PER>
<7,7,ORG>

system

<1,1,PER>
<5,5,PER>
<7,7,ORG>

LSTM with Keras

Entity linking

Michael	Jordan	can	dunk	from	the	free	throw	line
B-PER	I-PER							

Entity linking

- Task: Given a database of candidate referents, identify the correct referent for a mention in context.

Text	True wikipedia page
Hornets owner **Michael Jordan** thinks having one or two "superteams" is a detriment to the NBA because the other 28 teams "are going to be garbage."	`wiki/Michael_Jordan`
In 2001, **Michael Jordan** and others resigned from the Editorial Board of *Machine Learning*.	`wiki/Michael_I._Jordan`
The stars are aligning for leading man **Michael Jordan**, who just signed on for a new film, according to Variety.	`wiki/Michael_B._Jordan`
Michael Jordan played in 1,072 regular-season games in his 15-season career	`wiki/Michael_Jordan`

Wikification!

Michael Jordan (disambiguation)

From Wikipedia, the free encyclopedia

Michael Jordan (born 1963) is an American basketball player.

Michael or **Mike Jordan** may also refer to:

People [edit]

Sports [edit]

- Michael Jordan (footballer) (born 1986), English goalkeeper
- Mike Jordan (racing driver) (born 1958), English racing driver
- Mike Jordan (baseball, born 1863) (1863–1940), baseball player
- Mike Jordan (cornerback) (born 1992), American football cornerback
- Michael-Hakim Jordan (born 1977), American professional basketball player
- Michal Jordán (born 1990), Czech ice hockey player

Other people [edit]

- Michael B. Jordan (born 1987), American actor
- Michael Jordan (insolvency baron) (born 1931), English businessman
- Michael Jordan (Irish politician), Irish Farmers' Party TD from Wexford, 1927–1932
- Michael I. Jordan (born 1956), American researcher in machine learning and artificial intelligence
- Michael H. Jordan (1936–2010), American executive for CBS, PepsiCo, Westinghouse
- Michael Jordan (mycologist), English mycologist

High-level Algorithmic Approach

- Input: A text document d; Output: a set of pairs (m_i, t_i)
  - m_i are mentions in d; $t_j(m_i)$ are corresponding Wikipedia titles, or NIL.
- (1) Identify mentions $\mathbf{m_i}$ in d
- (2) Local Inference
  - For each $\mathbf{m_i}$ in d:
    - Identify a set of relevant titles $T(m_i)$
    - Rank titles $t_i \in T(m_i)$

    [E.g., consider local statistics of edges [(m_i, t_i), $(m_i, *)$, and $(*, t_i)$] occurrences in the Wikipedia graph]
- (3) Global Inference
  - For each document **d**:
    - Consider all $m_i \in$ d; and all $t_i \in T(m_i)$
    - Re-rank titles $t_i \in T(m_i)$

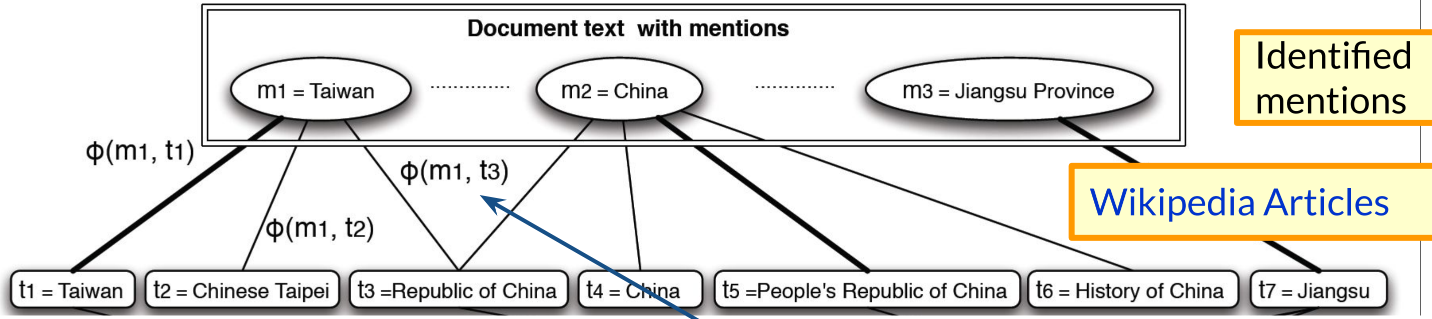
    [E.g., if m, m' are related by virtue of being in d, their corresponding titles t, t' may also be related]

Local approach

A text Document

Identified mentions

Wikipedia Articles



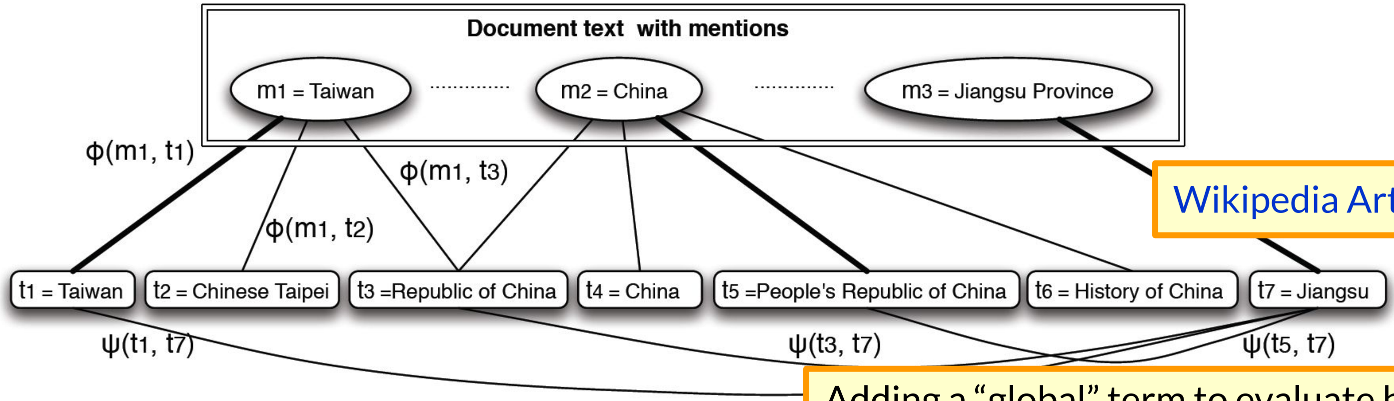
Local score of matching the mention to the title (decomposed by m_i)

- Γ is a solution to the problem
  - A set of pairs (m,t)
- m: a mention in the document
- t: the matched Wikipedia Title

$$\Gamma^*_{\text{local}} = \arg\max_{\Gamma} \sum_{i=1}^{N} \phi(m_i, t_i) \qquad (1)$$

Global Approach: Using Additional Structure

Text Document(s)—News, Blogs,...



Wikipedia Articles

$$\Gamma^* \approx \arg\max_{\Gamma} \sum_{i=1}^{N} [\phi(m_i, t_i) + \sum_{t_i \in \Gamma, t_j \in \Gamma'} \psi(t_i, t_j)]$$

Adding a "global" term to evaluate how good the structure of the solution is.

- Use the local solutions Γ' (each mention considered independently.
- Evaluate the structure based on pair-wise coherence scores $\Psi(t_i, t_j)$
- Choose those that satisfy document coherence conditions.

Candidate identification

NED Knowledge

41 million articles

294 languages

Rock music

Jeff Beck

From Wikipedia, the free encyclopedia

Geoffrey Arnold "Jeff" Beck (born 24 June 1944) is an English rock guitarist. He is one of the three noted guitarists to have played with The Yardbirds (the other two being Eric Clapton and Jimmy Page). Beck also formed The Jeff Beck Group and Beck, Bogert & Appice.

Much of Beck's recorded output has been instrumental, with a focus on innovative sound, and his releases have spanned genres ranging from blues rock, hard rock, jazz fusion, and an additional blend

Dictionary (candidate generation - text to article)

rock Rock_Music:90
the_yardbirds The_Yardbirds:467
eric_clapton Eric_Clapton:1098 Eric_Clapton_(album):78
…
beck Beck_Hansen:67 Jeff_Beck:3 Beck_Weathers:3 Beck_Mountain:1
jeff_beck Jeff_Beck:788 The_Jeff_Beck_Group:90

Learning to rank

- Entity linking is often cast as a learning to rank problem: given a mention x, some set of candidate entities $\mathcal{Y}$(x) for that mention, and context c, select the highest scoring entity from that set.

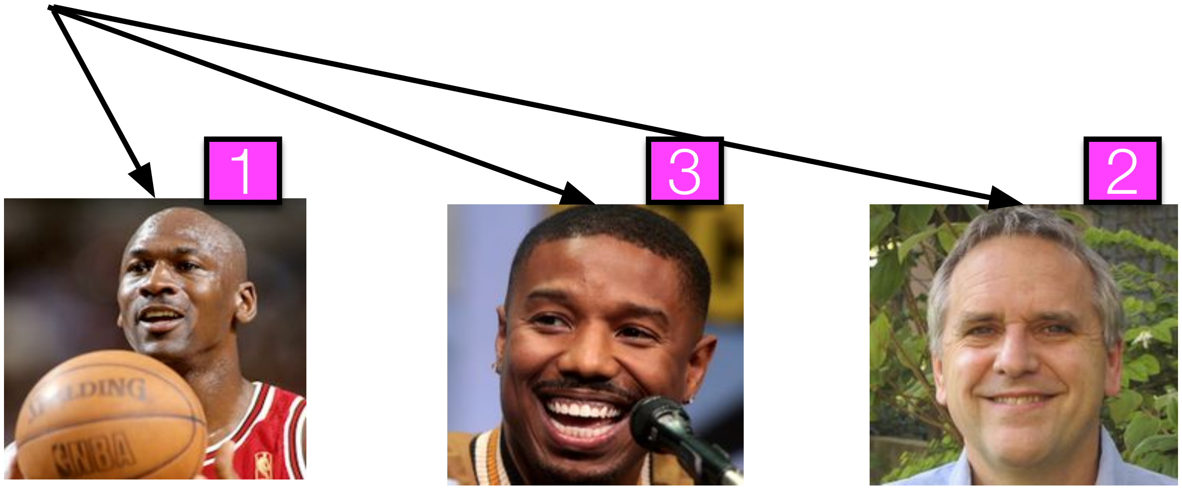
$$\hat{y} = \underset{y \in \mathcal{Y}(x)}{\arg\max}\ \Psi(y, x, c)$$

Some scoring function over the mention x, candidate y, and context c

Eisenstein 2018

Entity linking

Michael	Jordan	can	dunk	from	the	free	throw	line
B-PER	I-PER							



Learning to rank

- We learn the parameters of the scoring function by minimizing the **pairwise** ranking loss

$$\ell(\hat{y}, y, x, c) = \max\left(0, \Psi(\hat{y}, x, c) - \Psi(y, x, c) + 1\right)$$

other candidates

true candidate

Eisenstein 2018, Joachims 2002

Learning to rank

$$\ell(\hat{y}, y, x, c) = \max\left(0, \Psi(\hat{y}, x, c) - \Psi(y, x, c) + 1\right)$$

We suffer some loss if the predicted entity has a higher score than the true entity

$$\ell(\hat{y}, y, x, c) = \max\left(0, \Psi(\hat{y}, x, c) - \Psi(y, x, c) + 1\right)$$

You can't have a negative loss (if the true entity scores way higher than the predicted entity)

$$\ell(\hat{y}, y, x, c) = \max\left(0, \Psi(\hat{y}, x, c) - \Psi(y, x, c) + 1\right)$$

The true entity needs to score at least some constant margin better than the prediction; beyond that the higher score doesn't matter.

Learning to rank

Some scoring function over the mention x, candidate y, and context c

$$\Psi(y, x, c)$$

feature = f(x,y,c)
string similarity between x and y
popularity of y
NER type(x) = type(y)
cosine similarity between c and Wikipedia page for y

$$\Psi(y, x, c) = f(x, y, c)^{\top} \beta$$

Neural learning to rank

Parameters measuring the compatibility of the candidate and mention

Parameters measuring the compatibility of the candidate and context

$$\Psi(y, x, c) = v_y^\top \Theta^{(x,y)} x + v_y^\top \Theta^{(y,c)} c$$

Embedding for candidate

Embedding for mention

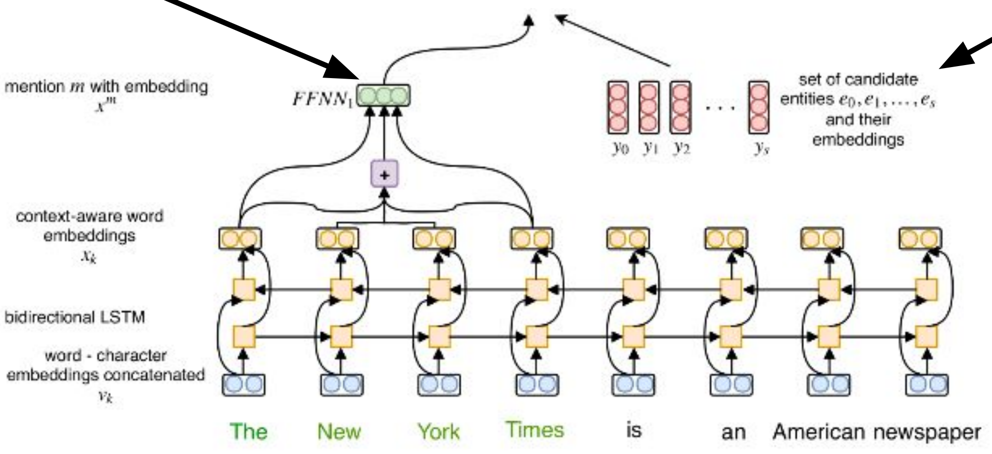
Embedding for context

Neural learning to rank

Embedding for mention

$$x, c) = v_y^{\top} \Theta^{(x,y)} x +$$

Embedding for candidate



Sil et al., 2018

Learning to rank

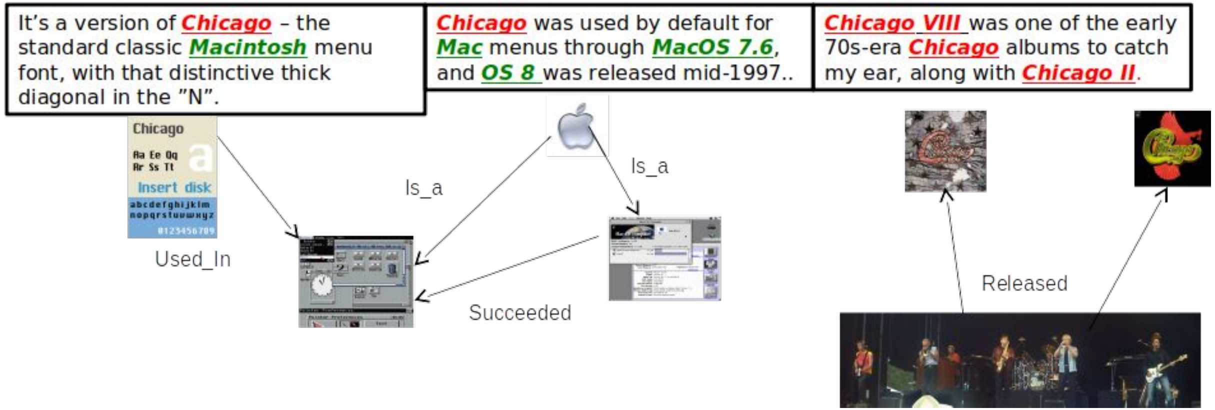
- We learn the parameters of the scoring by minimizing the ranking loss; take the derivative of the loss and backprop using SGD.

$$\ell(\hat{y}, y, x, c) = \max\left(0, \Psi(\hat{y}, x, c) - \Psi(y, x, c) + 1\right)$$

Eisenstein 2018

Collective entity linking

- **Recall**: The reference collection usually have a structure



- **Hypothesis**: Textual co-occurrences of concepts is reflected in KB (e.g. Wikipedia)
- **Incite**: Preferred linking contains structurally coherent concepts

Roth 2014,
Eisenstein 2018

Collective entity linking

$$\hat{\boldsymbol{y}} = \underset{\boldsymbol{y} \in \mathbb{Y}(\boldsymbol{x})}{\operatorname{argmax}} \Psi_c(\boldsymbol{y}) + \sum_{i=1}^{N} \Psi_\ell(y^{(i)}, \boldsymbol{x}^{(i)}, \boldsymbol{c}^{(i)})$$

The set of all possible collective entity assignments.

We can introduce a compatibility score over the set of entity assignment (global objective)

Same local scoring function over the mention x, candidate y, and context c

Eisenstein 2018

Collective entity linking

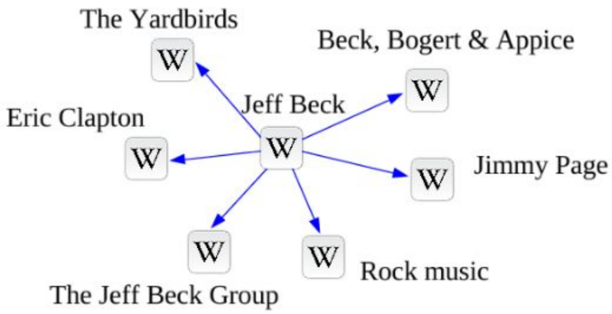
$$\hat{\boldsymbol{y}} = \underset{\boldsymbol{y} \in \mathbb{Y}(\boldsymbol{x})}{\operatorname{argmax}} \Psi_c(\boldsymbol{y}) + \sum_{i=1}^{N} \Psi_\ell(y^{(i)}, \boldsymbol{x}^{(i)}, \boldsymbol{c}^{(i)})$$

$$\Psi_c(\boldsymbol{y}) = \sum_{i=1}^{N} \sum_{j \neq i}^{N} \Psi_c(y^{(i)}, y^{(j)})$$

The compatibility score is typically reduced into a sum of pairwise scores

Eisenstein 2018

Collective entity linking



- Reward entity pairs for the number of Wikipedia categories they have in common (Cucerzan 2007)
- Number of incoming hyperlinks shared in the Wikipedia pages (Milne and Witten, 2008)
- Any graph based relatedness measures (e.g. PageRank) (Barrena et al., 2014)

$$\Psi_c(y^{(i)}, y^{(j)}) = \boldsymbol{v}_{y^{(i)}} \cdot \boldsymbol{v}_{y^{(j)}}$$

- Compatibility of two entities can be set as the similarity given by their embeddings

Eisenstein 2018

Lab session

- labs/3.NER_with_CRFs.ipynb