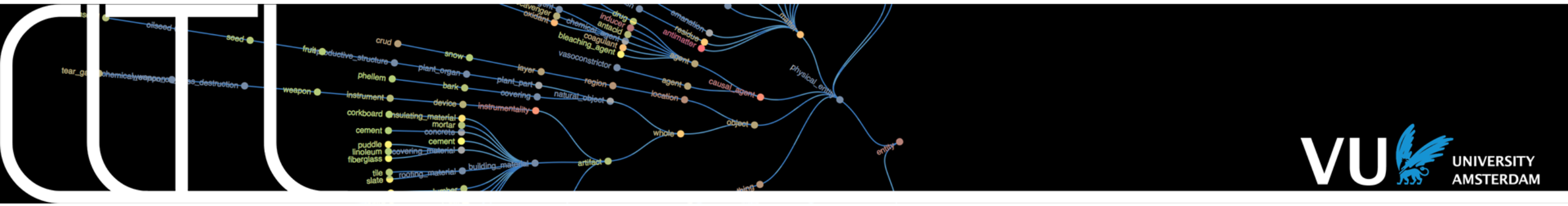
Text Mining CBS

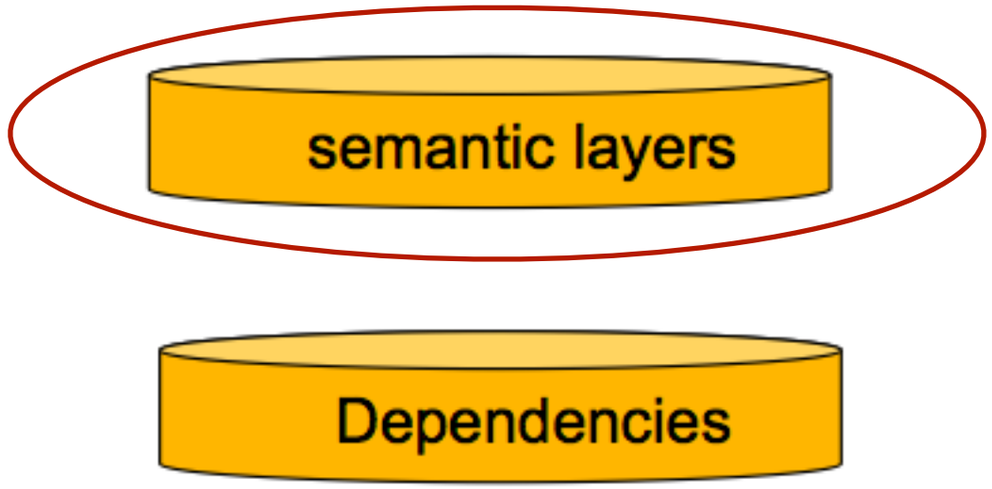
Lecture 4: Named entity detection and classification

Piek Vossen



Overview

- What is NER, NEC and NEL/NED?
- Approaches to NERC
- What is coreference of entity expressions, phrases and pronouns
- What is named entity disambiguation (NEL/NED)



What is an entity?

What is an entity?

What is reference?

What is an entity?

What is reference?

What is a referring expression?

What is an entity?
What is reference?
What is a referring expression?
What is a named entity expression?

What is Named Entity Recognition?

- Named Entity Recognition is the task of finding and classifying names in text.

Abraham Lincoln Listeni/ˈeɪbrəhæm ˈlɪŋkən/ (February 12, 1809 – April 15, 1865) was the 16th President of the United States, serving from March 1861 until his assassination in April 1865. Lincoln led the United States through its Civil War—its bloodiest war and its greatest moral, constitutional and political crisis.[1][2] In doing so, he preserved the Union, abolished slavery, strengthened the federal government, and modernized the economy.

Lincoln grew up on the western frontier in Kentucky and Indiana. Largely self-educated, he became a lawyer in Illinois, a Whig Party leader, and a member of the Illinois House of Representatives, where he served from 1834 to 1846. Elected to the United States House of Representatives in 1846, Lincoln promoted rapid modernization of the economy through banks, tariffs, and railroads. Because he had originally agreed not to run for a second term in Congress, and his opposition to the Mexican–American War was unpopular among Illinois voters, Lincoln returned to Springfield and resumed his successful law practice. Reentering politics in 1854, he became a leader in building the new Republican Party, which had a statewide majority in Illinois. In 1858, while taking part in a series of highly publicized debates with his opponent and rival, Democrat Stephen A. Douglas, Lincoln spoke out against the expansion of slavery, but lost the U.S. Senate race to Douglas.

source: http://en.wikipedia.org/wiki/Abraham_Lincoln

What is Named Entity Recognition?

- Named Entity Recognition is the task of finding and classifying names in text.

Abraham Lincoln Listeni/ˈeɪbrəhæm ˈlɪŋkən/ (February 12, 1809 – April 15, 1865) was the 16th President of the United States, serving from March 1861 until his assassination in April 1865. Lincoln led the United States through its Civil War—its bloodiest war and its greatest moral, constitutional and political crisis.[1][2] In doing so, he preserved the Union, abolished slavery, strengthened the federal government, and modernized the economy.

Lincoln grew up on the western frontier in Kentucky and Indiana. Largely self-educated, he became a lawyer in Illinois, a Whig Party leader, and a member of the Illinois House of Representatives, where he served from 1834 to 1846. Elected to the United States House of Representatives in 1846, Lincoln promoted rapid modernization of the economy through banks, tariffs, and railroads. Because he had originally agreed not to run for a second term in Congress, and his opposition to the Mexican–American War was unpopular among Illinois voters, Lincoln returned to Springfield and resumed his successful law practice. Reentering politics in 1854, he became a leader in building the new Republican Party, which had a statewide majority in Illinois. In 1858, while taking part in a series of highly publicized debates with his opponent and rival, Democrat Stephen A. Douglas, Lincoln spoke out against the expansion of slavery, but lost the U.S. Senate race to Douglas.

People

source: http://en.wikipedia.org/wiki/Abraham_Lincoln

What is Named Entity Recognition?

- Named Entity Recognition is the task of finding and classifying names in text.

Abraham Lincoln Listeni/ˈeɪbrəhæm ˈlɪŋkən/ (February 12, 1809 – April 15, 1865) was the 16th President of the United States, serving from March 1861 until his assassination in April 1865. Lincoln led the United States through its Civil War—its bloodiest war and its greatest moral, constitutional and political crisis.[1][2] In doing so, he preserved the Union, abolished slavery, strengthened the federal government, and modernized the economy.

Lincoln grew up on the western frontier in Kentucky and Indiana. Largely self-educated, he became a lawyer in Illinois, a Whig Party leader, and a member of the Illinois House of Representatives, where he served from 1834 to 1846. Elected to the United States House of Representatives in 1846, Lincoln promoted rapid modernization of the economy through banks, tariffs, and railroads. Because he had originally agreed not to run for a second term in Congress, and his opposition to the Mexican–American War was unpopular among Illinois voters, Lincoln returned to Springfield and resumed his successful law practice. Reentering politics in 1854, he became a leader in building the new Republican Party, which had a statewide majority in Illinois. In 1858, while taking part in a series of highly publicized debates with his opponent and rival, Democrat Stephen A. Douglas, Lincoln spoke out against the expansion of slavery, but lost the U.S. Senate race to Douglas.

People

Locations

source: http://en.wikipedia.org/wiki/Abraham_Lincoln

What is Named Entity Recognition?

- Named Entity Recognition is the task of finding and classifying names in text.

Abraham Lincoln Listeni/ˈeɪbrəhæm ˈlɪŋkən/ (February 12, 1809 – April 15, 1865) was the 16th President of the United States, serving from March 1861 until his assassination in April 1865. Lincoln led the United States through its Civil War—its bloodiest war and its greatest moral, constitutional and political crisis.[1][2] In doing so, he preserved the Union, abolished slavery, strengthened the federal government, and modernized the economy.

Lincoln grew up on the western frontier in Kentucky and Indiana. Largely self-educated, he became a lawyer in Illinois, a Whig Party leader, and a member of the Illinois House of Representatives, where he served from 1834 to 1846. Elected to the United States House of Representatives in 1846, Lincoln promoted rapid modernization of the economy through banks, tariffs, and railroads. Because he had originally agreed not to run for a second term in Congress, and his opposition to the Mexican–American War was unpopular among Illinois voters, Lincoln returned to Springfield and resumed his successful law practice. Reentering politics in 1854, he became a leader in building the new Republican Party, which had a statewide majority in Illinois. In 1858, while taking part in a series of highly publicized debates with his opponent and rival, Democrat Stephen A. Douglas, Lincoln spoke out against the expansion of slavery, but lost the U.S. Senate race to Douglas.

People

Locations

Organisations

source: http://en.wikipedia.org/wiki/Abraham_Lincoln

What is Named Entity Recognition?

- Named Entity Recognition is the task of finding and classifying names in text.

Abraham Lincoln Listeni/ˈeɪbrəhæm ˈlɪŋkən/ (February 12, 1809 – April 15, 1865) was the 16th President of the United States, serving from March 1861 until his assassination in April 1865. Lincoln led the United States through its Civil War—its bloodiest war and its greatest moral, constitutional and political crisis.[1][2] In doing so, he preserved the Union, abolished slavery, strengthened the federal government, and modernized the economy.

Lincoln grew up on the western frontier in Kentucky and Indiana. Largely self-educated, he became a lawyer in Illinois, a Whig Party leader, and a member of the Illinois House of Representatives, where he served from 1834 to 1846. Elected to the United States House of Representatives in 1846, Lincoln promoted rapid modernization of the economy through banks, tariffs, and railroads. Because he had originally agreed not to run for a second term in Congress, and his opposition to the Mexican–American War was unpopular among Illinois voters, Lincoln returned to Springfield and resumed his successful law practice. Reentering politics in 1854, he became a leader in building the new Republican Party, which had a statewide majority in Illinois. In 1858, while taking part in a series of highly publicized debates with his opponent and rival, Democrat Stephen A. Douglas, Lincoln spoke out against the expansion of slavery, but lost the U.S. Senate race to Douglas.

People

Time

Locations

Organisations

source: http://en.wikipedia.org/wiki/Abraham_Lincoln

What is Named Entity Recognition?

- Named Entity Recognition is the task of finding and classifying names in text.

Abraham Lincoln Listeni/ˈeɪbrəhæm ˈlɪŋkən/ (February 12, 1809 – April 15, 1865) was the 16th President of the United States, serving from March 1861 until his assassination in April 1865. Lincoln led the United States through its Civil War—its bloodiest war and its greatest moral, constitutional and political crisis.[1][2] In doing so, he preserved the Union, abolished slavery, strengthened the federal government, and modernized the economy.

Lincoln grew up on the western frontier in Kentucky and Indiana. Largely self-educated, he became a lawyer in Illinois, a Whig Party leader, and a member of the Illinois House of Representatives, where he served from 1834 to 1846. Elected to the United States House of Representatives in 1846, Lincoln promoted rapid modernization of the economy through banks, tariffs, and railroads. Because he had originally agreed not to run for a second term in Congress, and his opposition to the Mexican–American War was unpopular among Illinois voters, Lincoln returned to Springfield and resumed his successful law practice. Reentering politics in 1854, he became a leader in building the new Republican Party, which had a statewide majority in Illinois. In 1858, while taking part in a series of highly publicized debates with his opponent and rival, Democrat Stephen A. Douglas, Lincoln spoke out against the expansion of slavery, but lost the U.S. Senate race to Douglas.

People

Time

Locations

Events

Organisations

source: http://en.wikipedia.org/wiki/Abraham_Lincoln

Document Forensics
Network Institute project VU - Deloitte

- Investigating unsavoury business practices (e.g., slavery, fraud, bribery) can involve processing large numbers of contracts, yearly reports and external (news) sources that may reflect on a company's reputation and relations.

- Labour intensive task mainly using text search to identify relevant documents that are then manually processed.

- Project goals:

  - Extract the relevant concepts from unstructured texts (e.g. news) as well as semi-structured (e.g., contracts and financial) documents:

    - name of suppliers; the type of relationship between companies, executive management

  - Populate knowledge graphs and link them to publicly available knowledge graphs.

  - Knowledge graphs should reflect the temporal binding and provenance of the extracted relations and properties.

  - Enable automated reasoning about companies and their relationships such as structure of ownership or supply chains and their dynamics

Diligence detection

Auzina and Kim, 2019, Automated Due Diligence: Building Knowledge Graphs from News, Network Institute, VU University

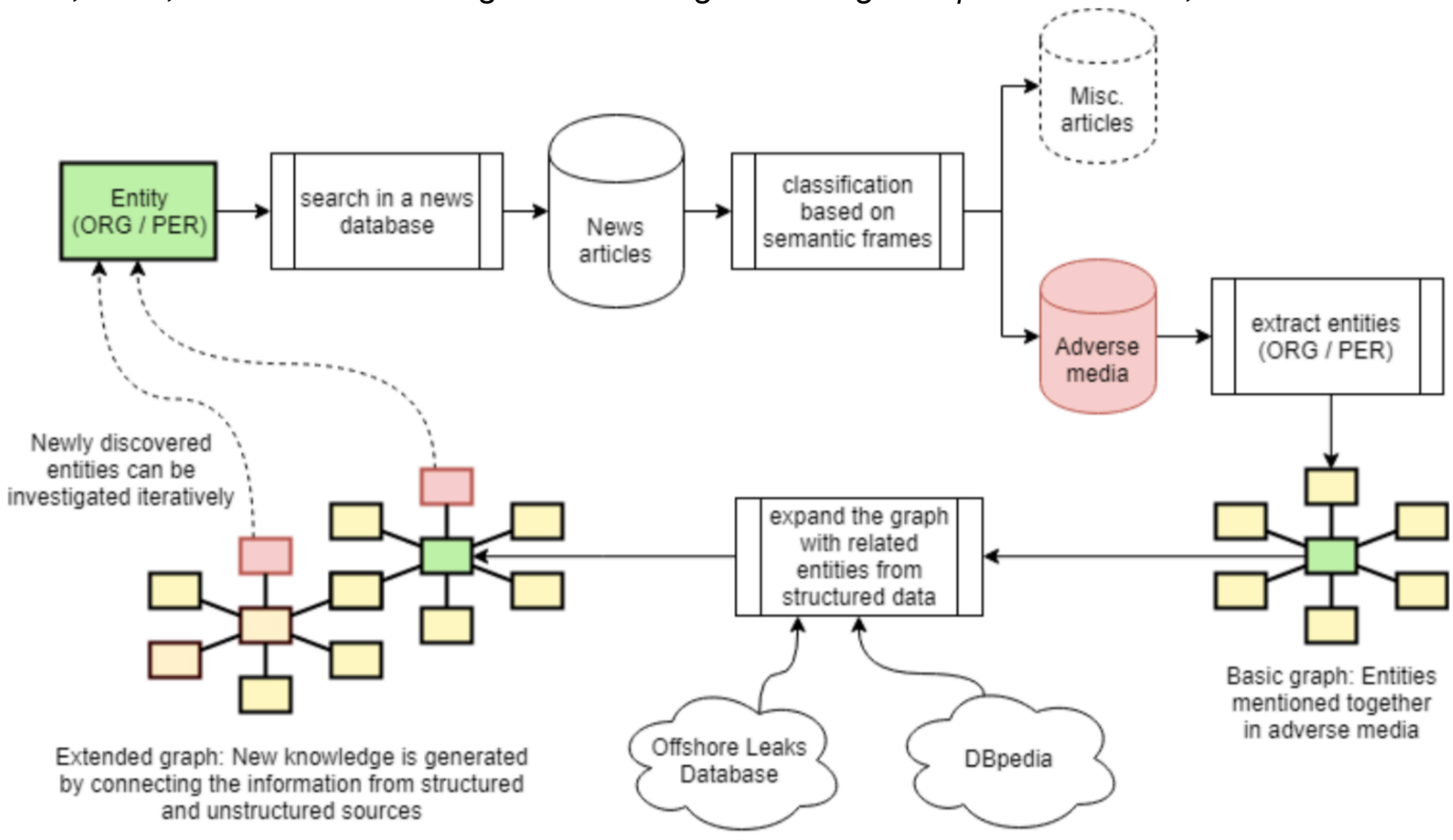


Figure 1: Overview of the proposed automated due diligence solution.

Adverse media classifier

Port of Moerdijk

Mitsubishi Materials

Kobe Steel

	MM dataset	KS dataset
Source	Nexis Uni[7]	Nexis Uni
Search term	Mitsubishi Materials	Kobe Steel
Content type	news	news
Language	English	English
Dates	06/91-02/19	01/00-04/19
# articles	707	1,774
# unique frames	460	540

Table 1: Datasets overview

Topic	# articles
MM dataset	
data falsification	65
forced labor during WWII	36
groundwater contamination	2
condos on contaminated soil	2
factory blast	1
KS dataset	
data falsification	115
tax evasion	2
asbestos-related employee death	1
employee embezzlement	1
safety and health violations	1

Table 2: Adverse media topics encountered during annotation

Model	Test set	Class	Precision	Recall	F1-score	Support
MM_10	KS: active learning sample (N=300)	0	0.65	0.89	0.75	177
		1	**0.67**	**0.31**	**0.42**	**123**
MM_10	KS: random sample (N=300)	0	0.90	0.95	0.92	242
		1	**0.73**	**0.55**	**0.63**	**58**
KS_10	MM: active learning sample (N=300)	0	0.82	0.91	0.86	194
		1	**0.80**	**0.63**	**0.71**	**106**
KS_10	MM: random sample (N=300)	0	0.91	0.97	0.94	240
		1	**0.82**	**0.62**	**0.70**	**60**

Table 3: Quantitative Evaluation Results (class 1: adverse media)

Auzina and Kim, 2019

Entity relationship graph

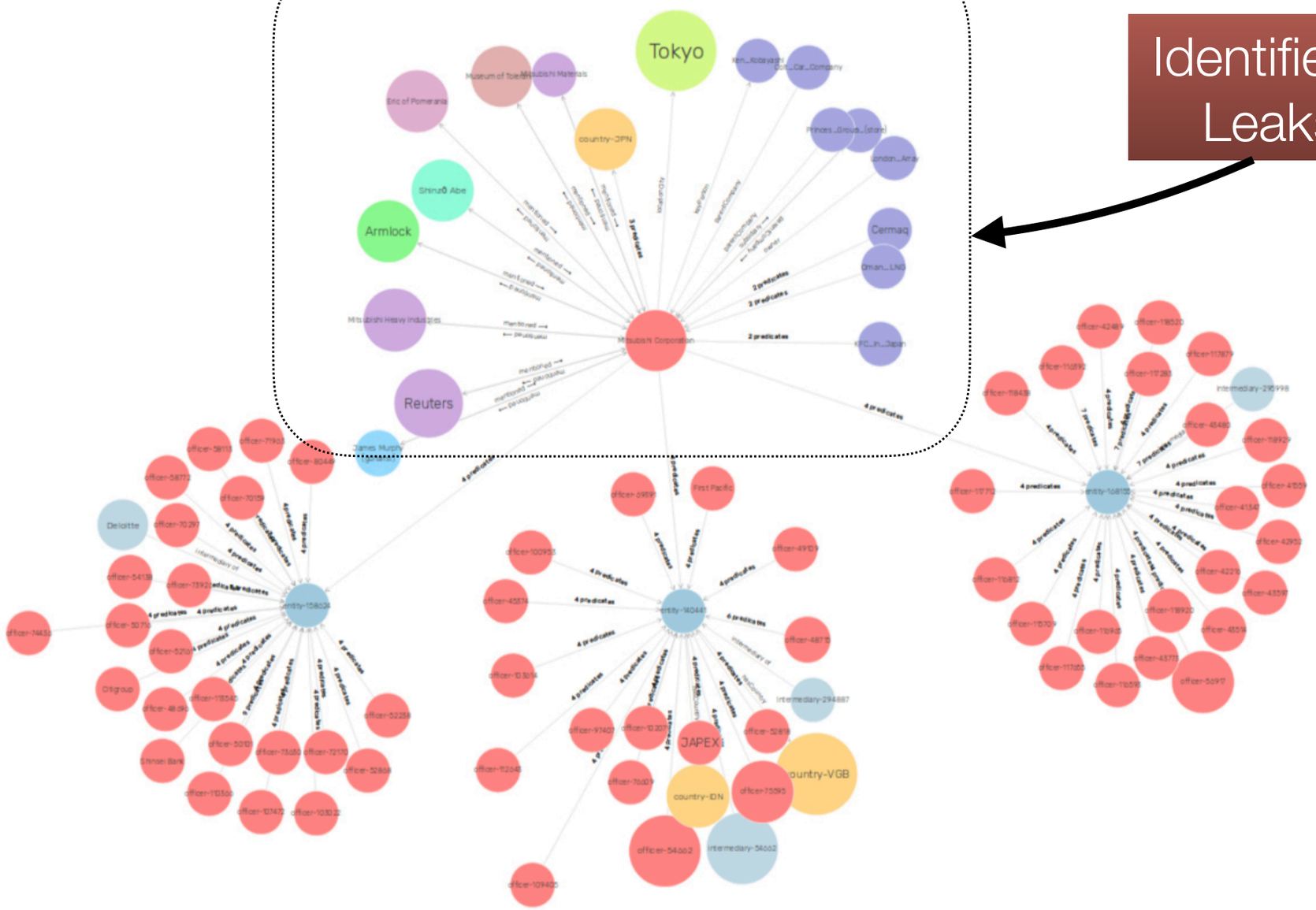


Figure 7: Extracted Officers and Intermediaries of the 3 identified entities

Auzina and Kim, 2019

Named entity detection and linking (NERC-D/L)

- NER(Recognition): detecting the phrase that is the name of an entity
- NEC(Classification): assigning an entity type to the phrase
- NEL(Linking) or NED(Disambiguation): establishing the identity of the entity in a given reference database (Wikipedia, DBPedia, YAGO)
- Coreference: any phrase that makes reference to an entity instance, including pronouns, noun phrases, abbreviations, acronyms, etc…
- Preprocessing: tokenisation, sentence splitting, Part-of-speech tagging, lookup, grammar rules, coreference

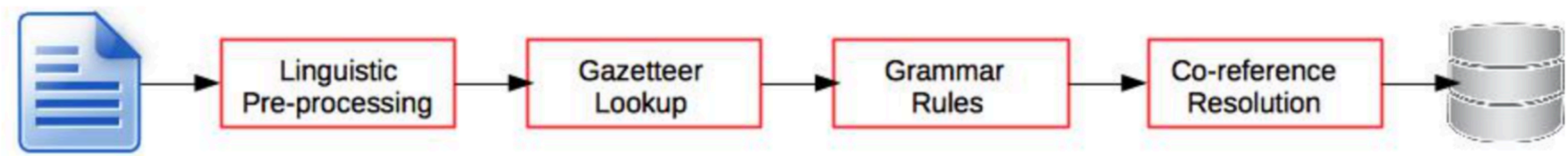


Figure 3.1: Typical NERC pipeline

What makes it a hard task?

- **variation** (IBM, The Big Blue, New York, NY, The Big Apple) and **ambiguity** **(**distinguish named entities and entities):
  - *MAY MAY RULE IN MAY*
  - *Austin Reed, Parkinson's disease, Pythagoras' Theorem*
- **extent**: *Sir Robert Walpole; [Abraham Lincoln, [the 16th President of the [United States]]]* <— nested entities
- **types**: e.g. *Criminal* as a subclass of *Person*, http://nerd.eurecom.fr/ontology, Fine-grained entity typing (e.g. FIGER uses 112 types from Freebase)
- **Time**: 8am, yesterday, last week, this month/year (TimeML types DAY, TIME, DURATION, SET)
- **Metonymy**: US, Holland, The Netherlands, Ford, Volkswagen

NERC feature engineering

- Word level features
- Digit patterns
- Common word endings
- Functions over words: non-alphabetic (A.T.&T.), n-grams
- Lookup features
- Document & Corpus features

Word level features

Table 1. Word-level features

Features	Examples
Case	– Starts with a capital letter – Word is all uppercased – The word is mixed case (e.g., ProSys, eBay)
Punctuation	– Ends with period, has internal period (e.g., St., I.B.M.) – Internal apostrophe, hyphen or ampersand (e.g., O'Connor)
Digit	– Digit pattern (*see Section 3.1.1*) – Cardinal and ordinal – Roman number – Word with digits (e.g., W3C, 3M)
Character	– Possessive mark, first person pronoun – Greek letters
Morphology	– Prefix, suffix, singular version, stem – Common ending (*see Section 3.1.2*)
Part-of-speech	– proper name, verb, noun, foreign word
Function	– Alpha, non-alpha, n-gram (*see Section 3.1.3*) – lowercase, uppercase version – pattern, summarized pattern (*see Section 3.1.4*) – token length, phrase length

From Nadeau, D., & Sekine, S. (2007)

Gazetteers & lexicons

Table 2. List lookup features.

Features	Examples
General list	– General dictionary (see Section 3.2.1) – Stop words (function words) – Capitalized nouns (e.g., January, Monday) – Common abbreviations
List of entities	– Organization, government, airline, educational – First name, last name, celebrity – Astral body, continent, country, state, city
List of entity cues	– Typical words in organization (see 3.2.2) – Person title, name prefix, post-nominal letters – Location typical word, cardinal point

From Nadeau, D., & Sekine, S. (2007)

Document features

Table 3. Features from documents.

Features	Examples
Multiple occurrences	– Other entities in the context – Uppercased and lowercased occurrences (see 3.3.1) – Anaphora, coreference (see 3.3.2)
Local syntax	– Enumeration, apposition – Position in sentence, in paragraph, and in document
Meta information	– Uri, email header, XML section, (see Section 3.3.3) – Bulleted/numbered lists, tables, figures
Corpus frequency	– Word and phrase frequency – Co-occurrences – Multiword unit permanency (see 3.3.4)

From Nadeau, D., & Sekine, S. (2007)

CONLL: Computational Natural Language Learning
https://www.conll.org

Every token on a separate line followed by TAB separated columns with annotations (TSV)

ID	FORM	LEMMA	UPOSTAG	XPOSTAG	FEATS	HEAD	DEPREL	DEPS	NER
# newdoc url = http://www.poweredbyosteons.org/2012/01/brief-history-of-bioarchaeological.html									
# newdoc s3 = s3://aws-publicdatasets/common-crawl/crawl-data/CC-MAIN-2016-07/segments...									
...									
# sent_id = http://www.poweredbyosteons.org/2012/01/brief-history-of-bioarchaeological.html#60									
# text = The American Museum of Natural History was established in New York in 1869.									
0	The	the	DT	DT	_	2	det	2:det	O
1	American	American	NNP	NNP	_	2	nn	2:nn	B-Organization
2	Museum	Museum	NNP	NNP	_	7	nsubjpass	7:nsubjpass	I-Organization
3	of	of	IN	IN	_	2	prep	_	I-Organization
4	Natural	Natural	NNP	NNP	_	5	nn	5:nn	I-Organization
5	History	History	NNP	NNP	_	3	pobj	2:prep_of	I-Organization
6	was	be	VBD	VBD	_	7	auxpass	7:auxpass	O
7	established	establish	VBN	VBN	_	7	ROOT	7:ROOT	O
8	in	in	IN	IN	_	7	prep	_	O
9	New	New	NNP	NNP	_	10	nn	10:nn	B-Location
10	York	York	NNP	NNP	_	8	pobj	7:prep_in	I-Location
11	in	in	IN	IN	_	7	prep	_	O
12	1869	1869	CD	CD	_	11	pobj	7:prep_in	O
13	.	.	.	.	_	7	punct	7:punct	O
...									

https://www.clips.uantwerpen.be/conll2003/ner/

IO(B) style
I = insight
O = outside
B = beginning

CoNLL feature representation

Word	POS	Chunk	Short shape	Label
American	NNP	B-NP	Xx	B-ORG
Airlines	NNPS	I-NP	Xx	I-ORG
,	,	O	,	O
a	DT	B-NP	x	O
unit	NN	I-NP	x	O
of	IN	B-PP	x	O
AMR	NNP	B-NP	X	B-ORG
Corp.	NNP	I-NP	Xx.	I-ORG
,	,	O	,	O
immediately	RB	B-ADVP	x	O
matched	VBD	B-VP	x	O
the	DT	B-NP	x	O
move	NN	I-NP	x	O
,	,	O	,	O
spokesman	NN	B-NP	x	O
Tim	NNP	I-NP	Xx	B-PER
Wagner	NNP	I-NP	Xx	I-PER
said	VBD	B-VP	x	O
.	,	O	.	O

Sequence labelling

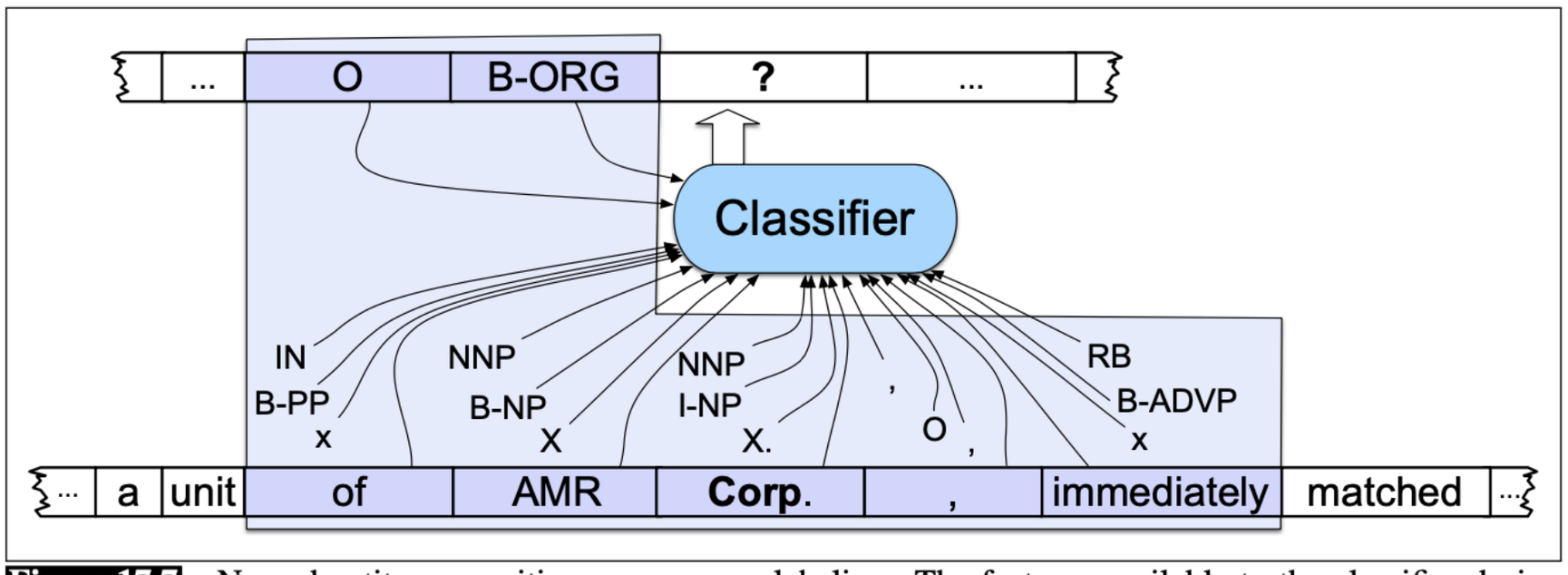
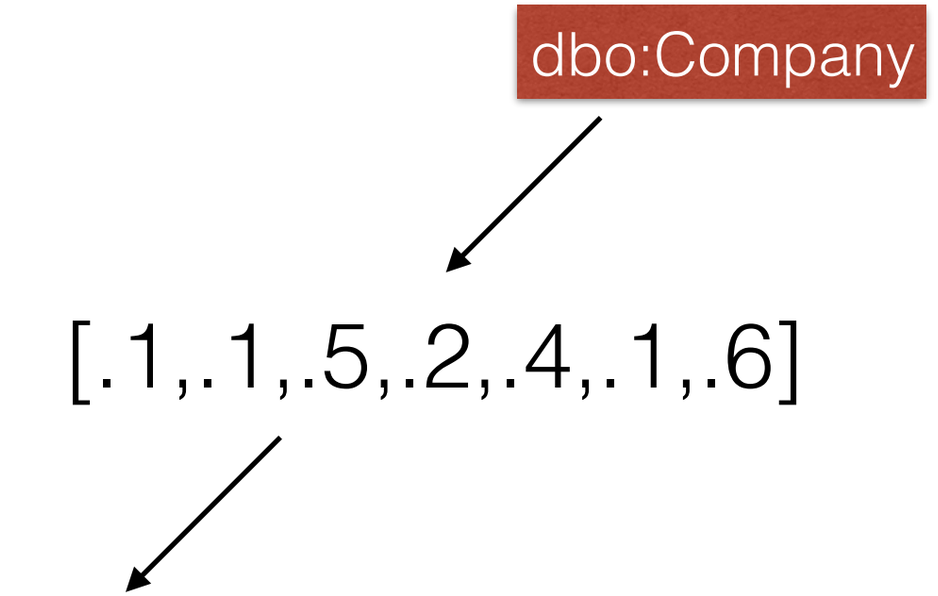


Figure 17.7 Named entity recognition as sequence labeling. The features available to the classifier during training and classification are those in the boxed area.

Juravsky & Martin 3ed, 2018

Entity embeddings

- **Light entity:** Groupon (dbo:Company)
  - **Met de korting**sbonnen van Groupon kunt u tegen hoge **korting kennismaken** met de diensten van een professionele fotograaf.
  - **Met de** Groupon **app** kan je nu ook onderweg de deals bekijken, kopen en inwisselen.
- **Dark Entity**: Scoupy
  - Via de gratis **app** voor iPhone of Android of via de website www.scoupy.nl kun je met **hoge korting** of zelfs helemaal gratis **kennismaken** met allerlei producten.
  - **Met de** Scoupy **app** kan je op zoek naar **korting**scoupons voor winkels en restaurants bij jou in de buurt.



[.1,.1,.5,.2,.4,.1,.6]

- Groupon
  [.3,.1,.2,.3,.7,.2,.3]

  Scoupy
  [.2,.1,.3,.2,.5,.1,.1]

Word embeddings

300-500 dimensions

Converting features to one-hot-vectors & embeddings

- "The president of Groupon eats an apple"
  - Groupon $[1]_{Case}$ +$[7]_{Length}$ + $[0,0,1,0,0,0]_{PoS}$ +

    $[0,0,00,0,0,0,0,0,0,0,1]_{Word}$ + $[1]_{Gazetteer}$ + $[.1,.3,.2,.6,.7.,2]_{Embedding}$
- "A manager of Scoupy swallows the banana"
  - Scoupy $[1]_{Case}$ + $[6]_{Length}$ + $[0,0,1,0,0,0]_{PoS}$ +

    $[0,0,00,0,0,0,0,0,0,1,0]_{Word}$ + $[0]_{Gazetteer}$ +$[.1,.2,.3,.4,.1.,1]_{Embedding}$

NERC as a sequence tagging task

- Sentences exhibit strong predictive probabilities for sequences of words and their tags, including IOB entity tags:
    - Abraham (B-PER) Lincoln (I-PER) ((O) February (B-T) 12 (I-T) , (I-T) 1809 (I-T)
- CRFs (Conditional Random Fields) are one of the most widely used algorithms for NERC
    - Graph models view NERC as a sequence classification task
    - Strong dependence between features and predictions in a sequence, e.g. I-LOC never occurs immediately after B-PER
- https://www.quora.com/What-are-the-pros-and-cons-of-these-three-sequence-models-MaxEnt-Markov-Model-Conditional-random-fields-and-recurrent-neural-networks

Sequence tagging problems in NLP
Part of speech tagging using Hidden Markov Models (HMM)

- http://www.cse.unsw.edu.au/~billw/cs9414/notes/nlp/ambiguity/ambiguity-2009.html

Large collection of text annotated with Part-of-Speech tags

[I, do, not, like, flies, on, my, food]
[Pr, V, Av, V, N, P, PR, N]
[She, flies, to, China, after, the, meeting]
[Pr, V, P, N, P, ART, N]

Table 7.6 The lexical generation probabilities

Pr(the | ART)	0.54		Pr(a | ART)	0.360
Pr(flies | N)	0.025		Pr(a | N)	0.001
Pr(flies | V)	0.076		Pr(flower | N)	0.063
Pr(like | V)	0.1		Pr(flower | V)	0.05
Pr(like | P)	0.068		Pr(birds | N)	0.076
Pr(like | N)	0.012			

Sequence tagging problems in NLP
Part of speech tagging using Hidden Markov Models (HMM)

- http://www.cse.unsw.edu.au/~billw/cs9414/notes/nlp/ambiguity/ambiguity-2009.html

Viterbi algorithm

[flies, like, a, flower]

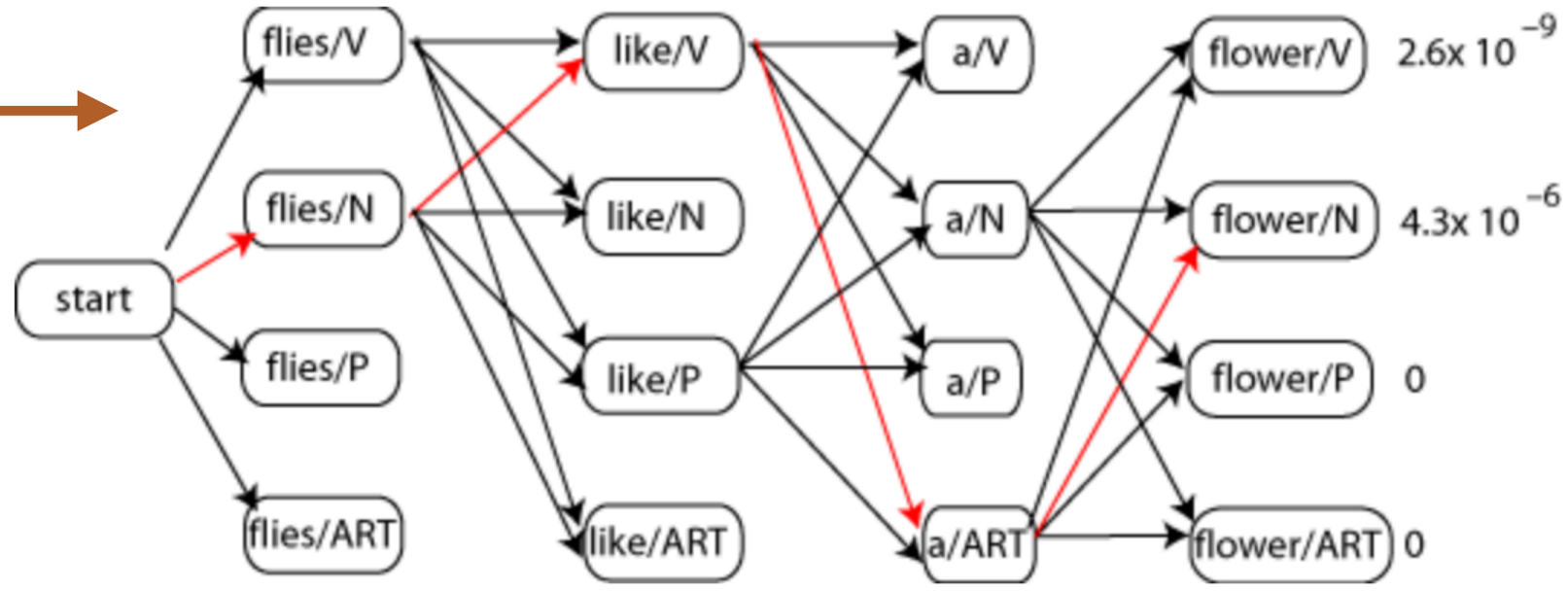
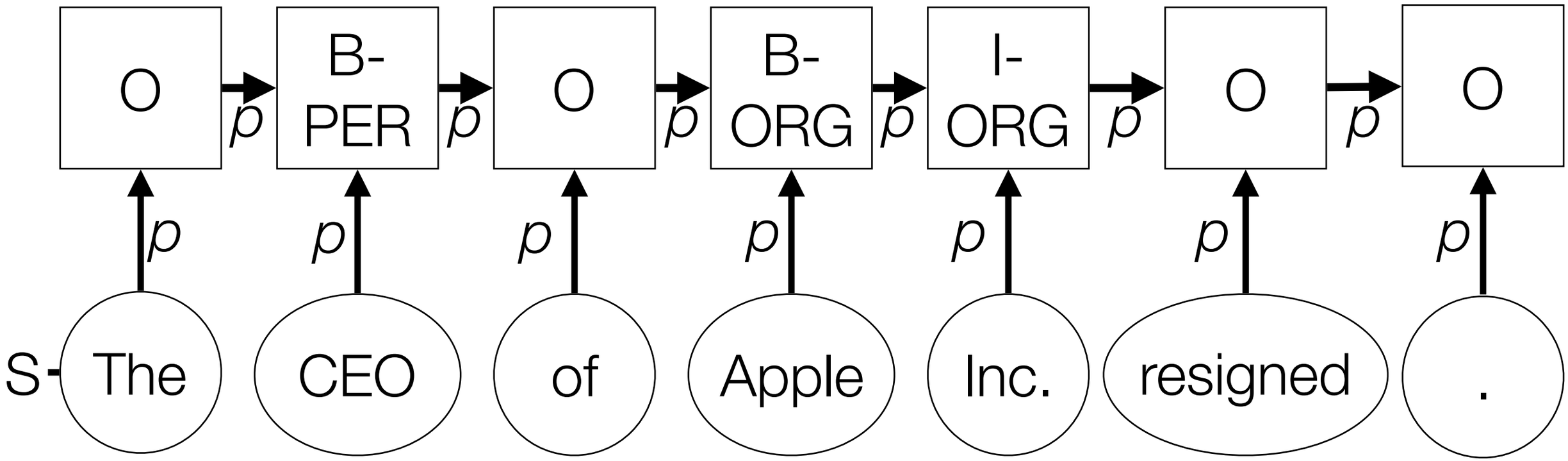


Table 7.6 The lexical generation probabilitie

Pr(the | ART)	0.54		Pr(a | ART)	0.3
Pr(flies | N)	0.025		Pr(a | N)	0.0
Pr(flies | V)	0.076		Pr(flower | N)	0.0
Pr(like | V)	0.1		Pr(flower | V)	0.0
Pr(like | P)	0.068		Pr(birds | N)	0.0
Pr(like | N)	0.012			

lexcat	SeqScore (lexcat,1)	SeqScore (lexcat,2)	SeqScore (lexcat,3)	SeqScore (lexcat,4)	BackPtr (lexcat,4)
V	7.6×10^{-6}	0.00031	0	2.6×10^{-9}	ART
N	0.00725	1.3×10^{-5}	1.2×10^{-7}	4.3×10^{-6}	ART
P	0	0.00022	0	0	∅
ART	0	0	7.2×10^{-5}	0	∅

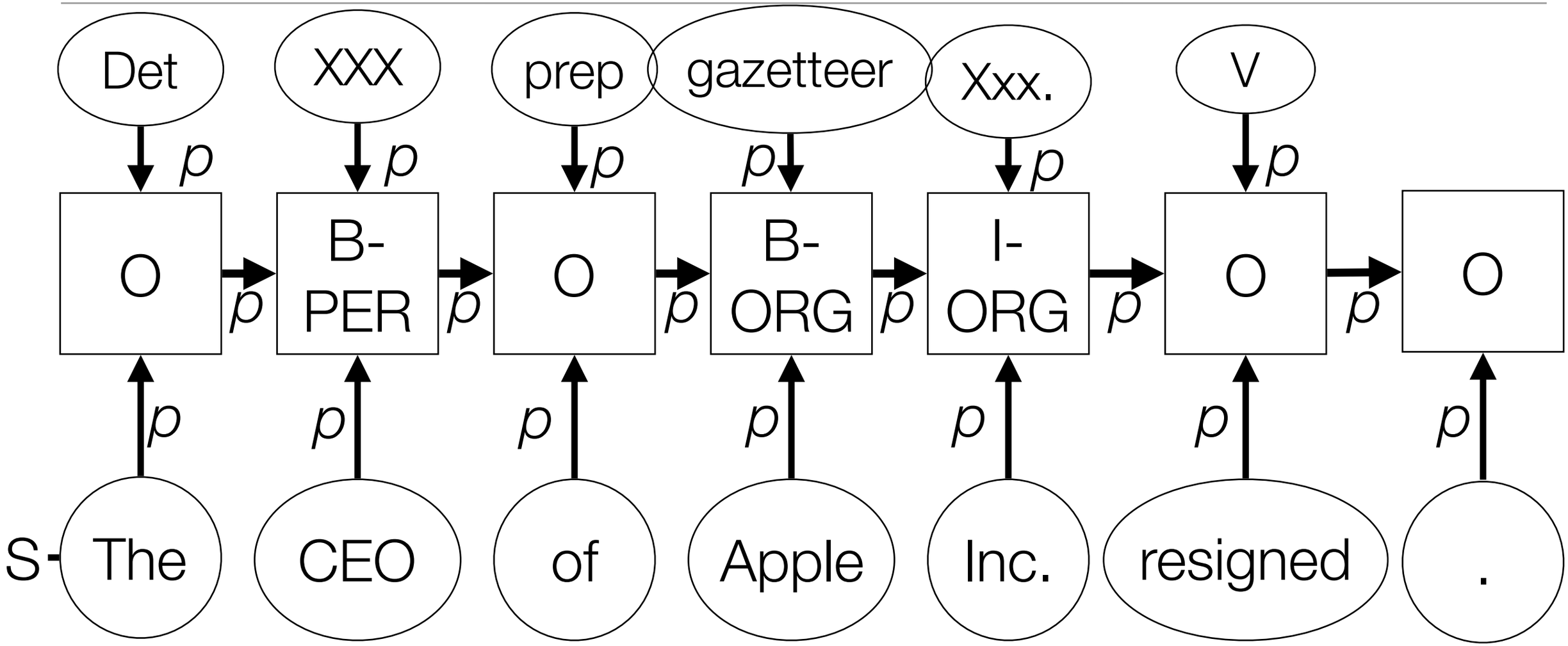
Conditional Random Field for IOB sequences



condition a: if *all caps* & preceded by O then B

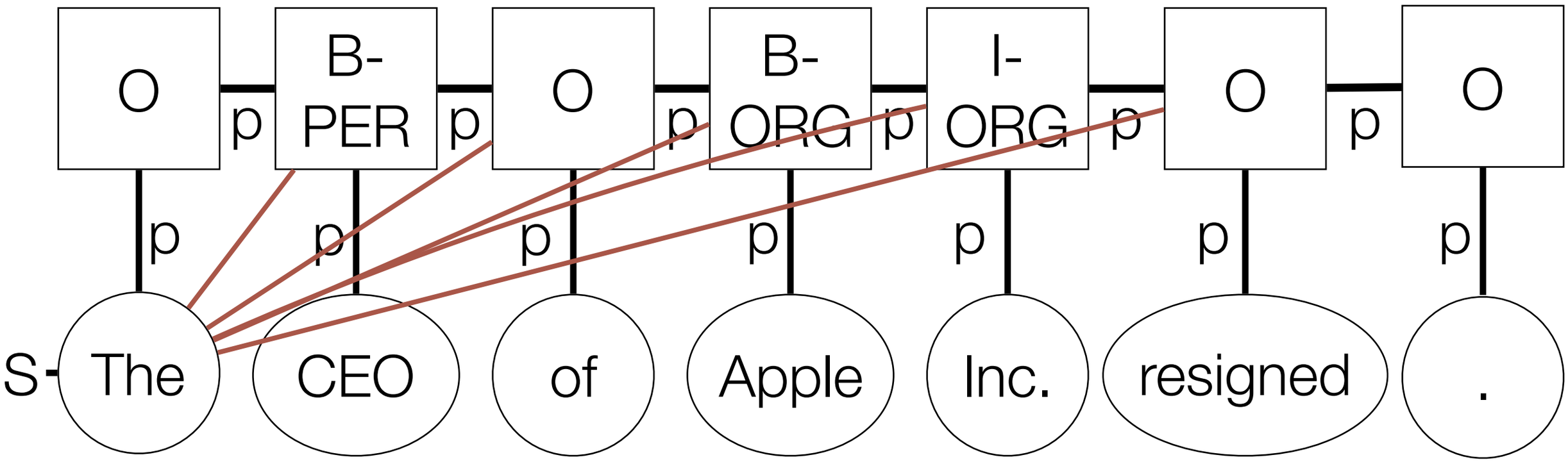
condition b: if *initial capital* & followed by “Inc.” then ORG

Conditional Random Field for IOB sequences

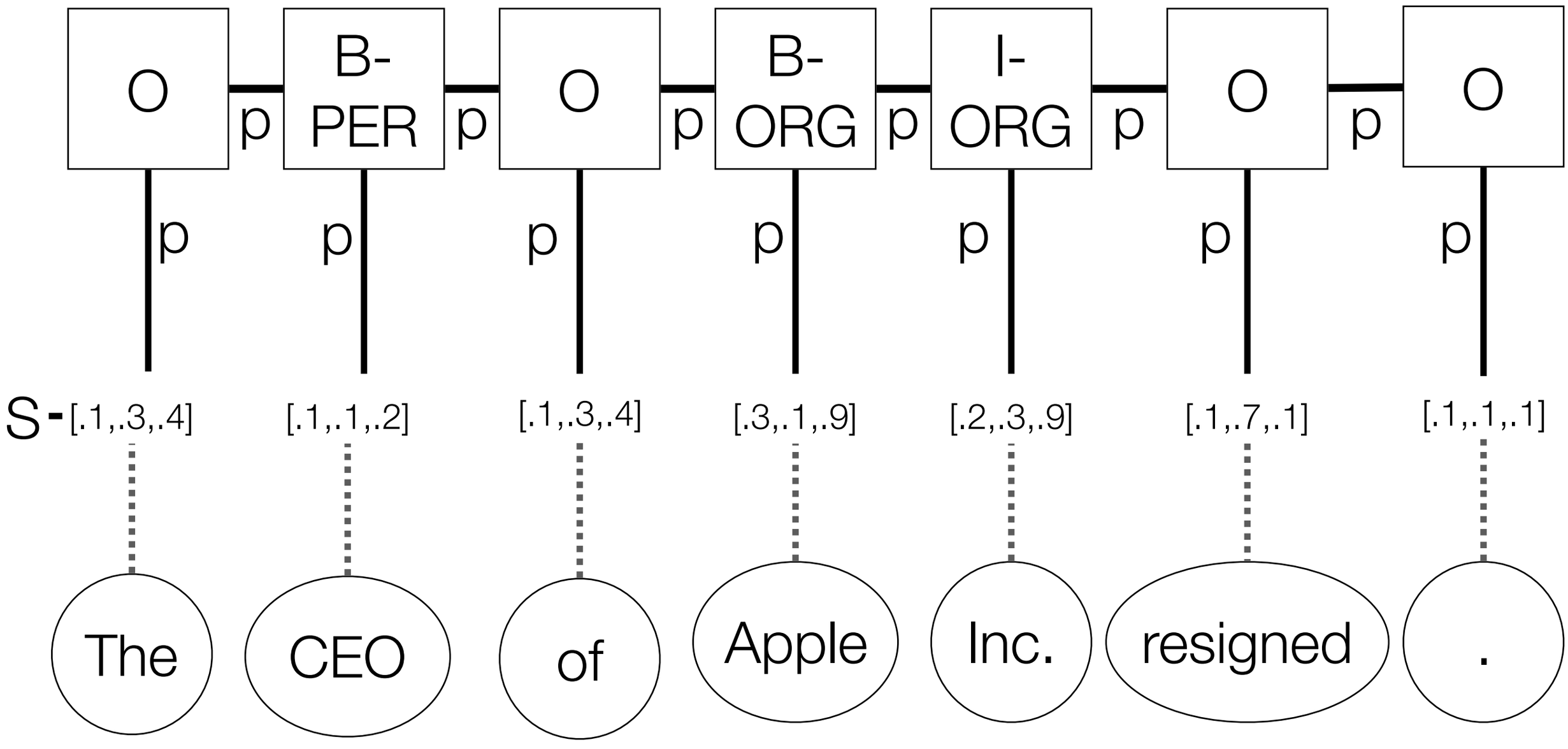


- Learn predictive probabilities of many features and sequential dependencies (w+1, w+2, w+3, etc.) to predict labels: I, O, B or I-PER, I-PER, B-PER, I-ORG, I-ORG, B-ORG, etc.

Conditional Random Field for IOB sequences



Conditional Random Field for IOB sequences



NERC performance

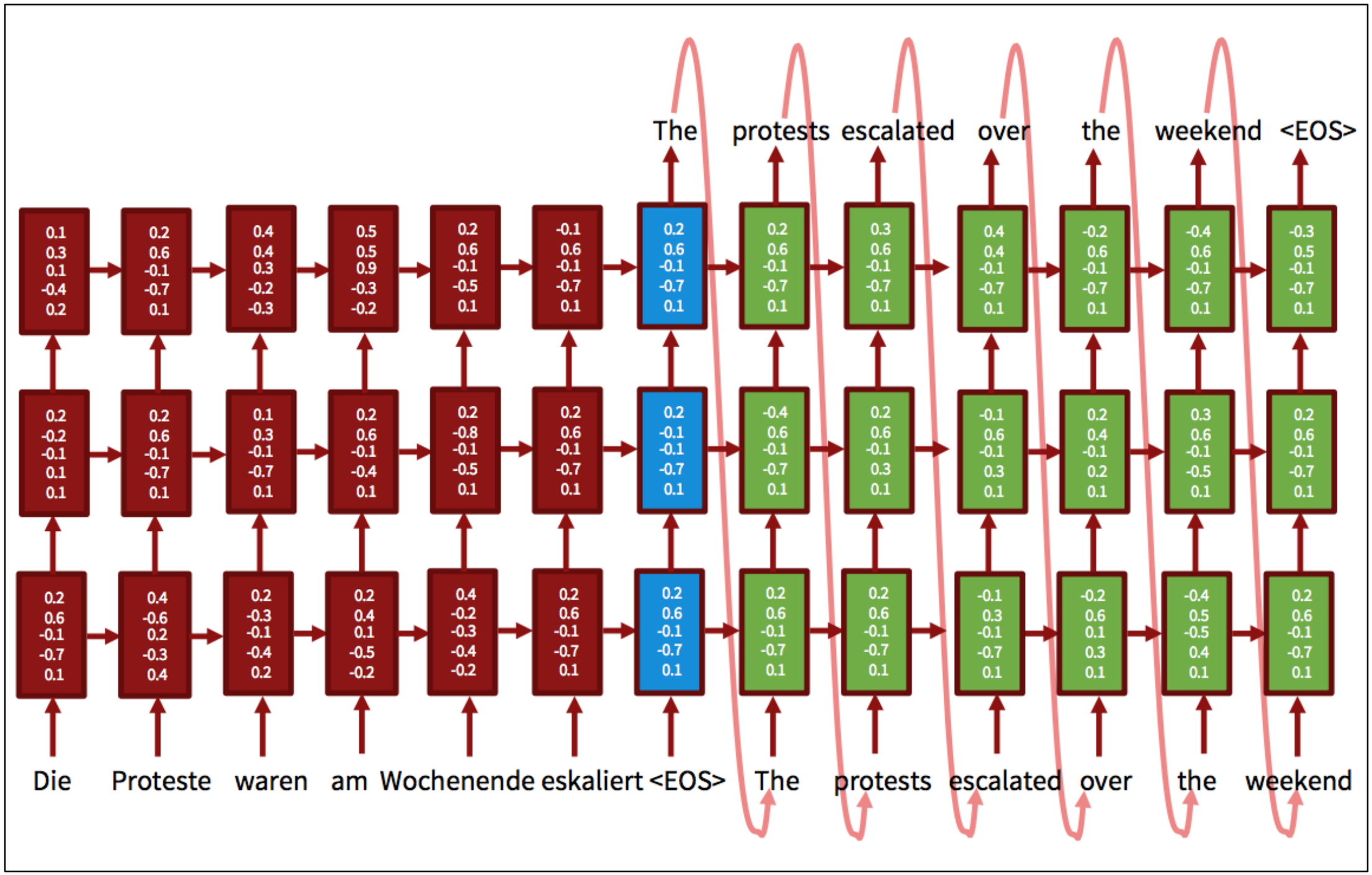
Feature-engineered machine learning systems	Dict	SP	DU	EN	GE
Carreras et al. (2002) binary AdaBoost classifiers	Yes	81.39	77.05	-	-
Malouf (2002) - Maximum Entropy (ME) + features	Yes	73.66	68.08	-	-
Li et al. (2005) SVM with class weights	Yes	-	-	88.3	-
Passos et al. (2014) CRF	Yes	-	-	90.90	-
Ando and Zhang (2005a) Semi-supervised state of the art	No	-	-	89.31	75.27
Agerri and Rigau (2016)	Yes	**84.16**	**85.04**	**91.36**	**76.42**
Feature-inferring neural network word models					
Collobert et al. (2011) Vanilla NN +SLL / Conv-CRF	No	-	-	81.47	-
Huang et al. (2015) Bi-LSTM+CRF	No	-	-	84.26	-
Yan et al. (2016) Win-BiLSTM (English), FF (German) (Many fets)	Yes	-	-	88.91	**76.12**
Collobert et al. (2011) Conv-CRF (SENNA+Gazetteer)	Yes	-	-	89.59	-
Huang et al. (2015) Bi-LSTM+CRF+ (SENNA+Gazetteer)	Yes	-	-	**90.10**	-
Feature-inferring neural network character models					
Gillick et al. (2015) – BTS	No	**82.95**	**82.84**	**86.50**	**76.22**
Kuru et al. (2016) CharNER	No	82.18	79.36	84.52	70.12
Feature-inferring neural network word + character models					
Yang et al. (2017)	Yes	85.77	**85.19**	91.26	-
Luo (2015)	Yes	-	-	91.20	-
Chiu and Nichols (2015)	Yes	-	-	**91.62**	-
Ma and Hovy (2016)	No	-	-	91.21	-
Santos and Guimaraes (2015)	No	82.21	-	-	-
Lample et al. (2016)	No	85.75	81.74	90.94	**78.76**
Bharadwaj et al. (2016)	Yes	**85.81**	-	-	-
Dernoncourt et al. (2017)	No	-	-	90.5	-
Feature-inferring neural network word + character + affix models					
Re-implementation of Lample et al. (2016) (100 Epochs)	No	85.34	85.27	90.24	78.44
Yadav et al. (2018)(100 Epochs)	No	86.92	87.50	90.69	78.56
Yadav et al. (2018) (150 Epochs)	No	**87.26**	**87.54**	90.86	**79.01**

Table 1: Comparison of NER systemsin four languages: CoNLL 2002 Spanish (SP), CoNLL 2002 Dutch (DU), CoNLL 2003 English (EN), and CoNLL 2003 German (GE). Dict indicates whether or not the approach makes use of dictionary lookups. Best performance in each category is highlighted in bold.

Yadav, Vikas, and Steven Bethard. "A survey on recent advances in named entity recognition from deep learning models." In *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 2145-2158. 2018.

Feature-inferring NN systems outperform feature-engineered systems, despite the latter's access to domain specific rules, knowledge, features, and lexicons

Machine translation: long-short-term-memory



Google Neural MT

Neural network (LSTM) and CRF

https://github.com/guillaumegenthial/tf_ner

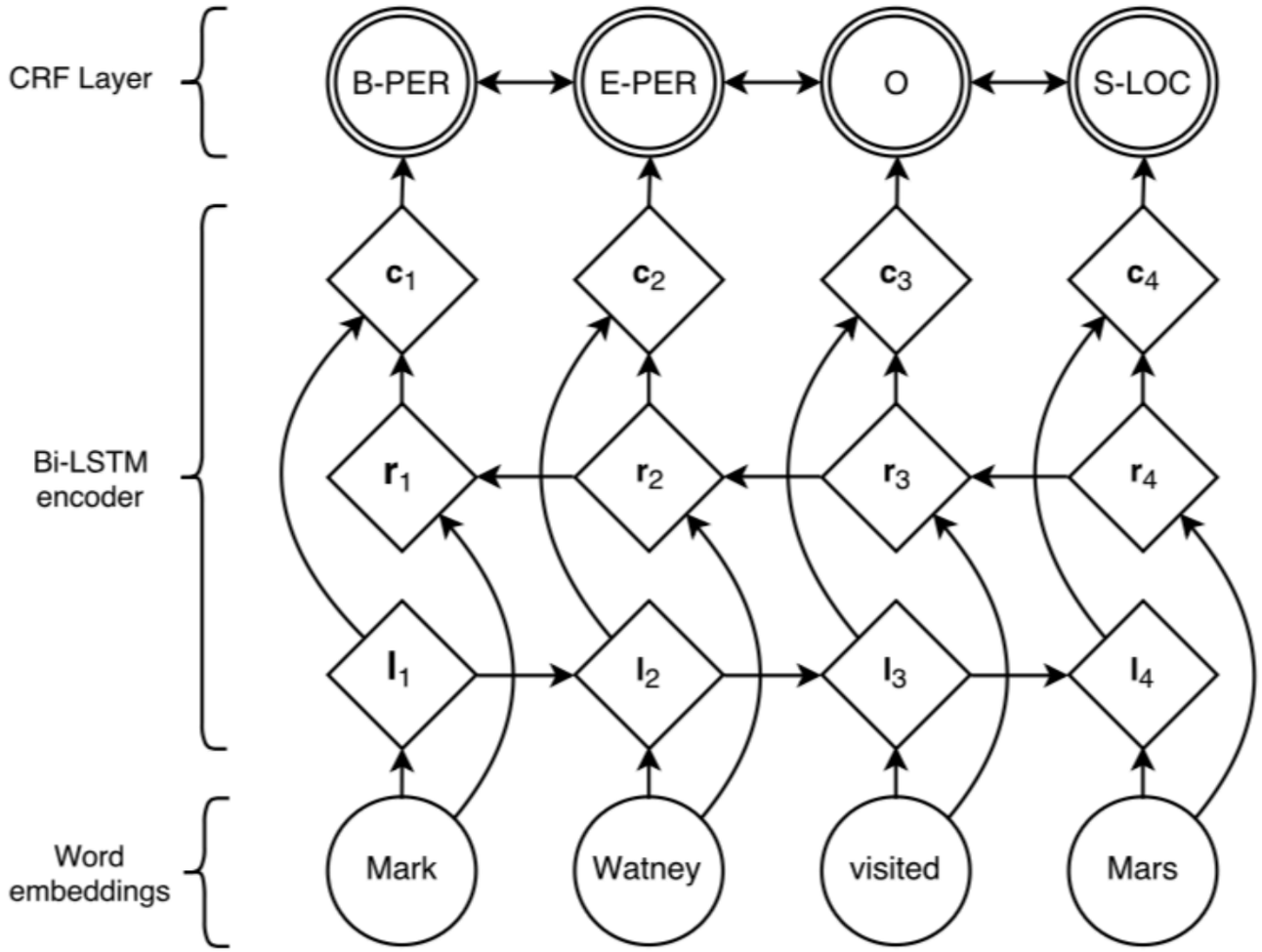


Figure 1: Main architecture of the network. Word embeddings are given to a bidirectional LSTM. $\mathbf{l}_i$ represents the word i and its left context, $\mathbf{r}_i$ represents the word i and its right context. Concatenating these two vectors yields a representation of the word i in its context, $\mathbf{c}_i$.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami and Chris Dyer, 2016, Neural Architectures for Named Entity Recognition, NAACL.

Long Short-Term Memory LSTM

Zhiheng Huang, Wei Xu, Kai Yu 2015, Bidirectional LSTM-CRF Models for Sequence Tagging, arXiv.1508.01991v1

CRF on top of bi-LSTM
using word & character embeddings

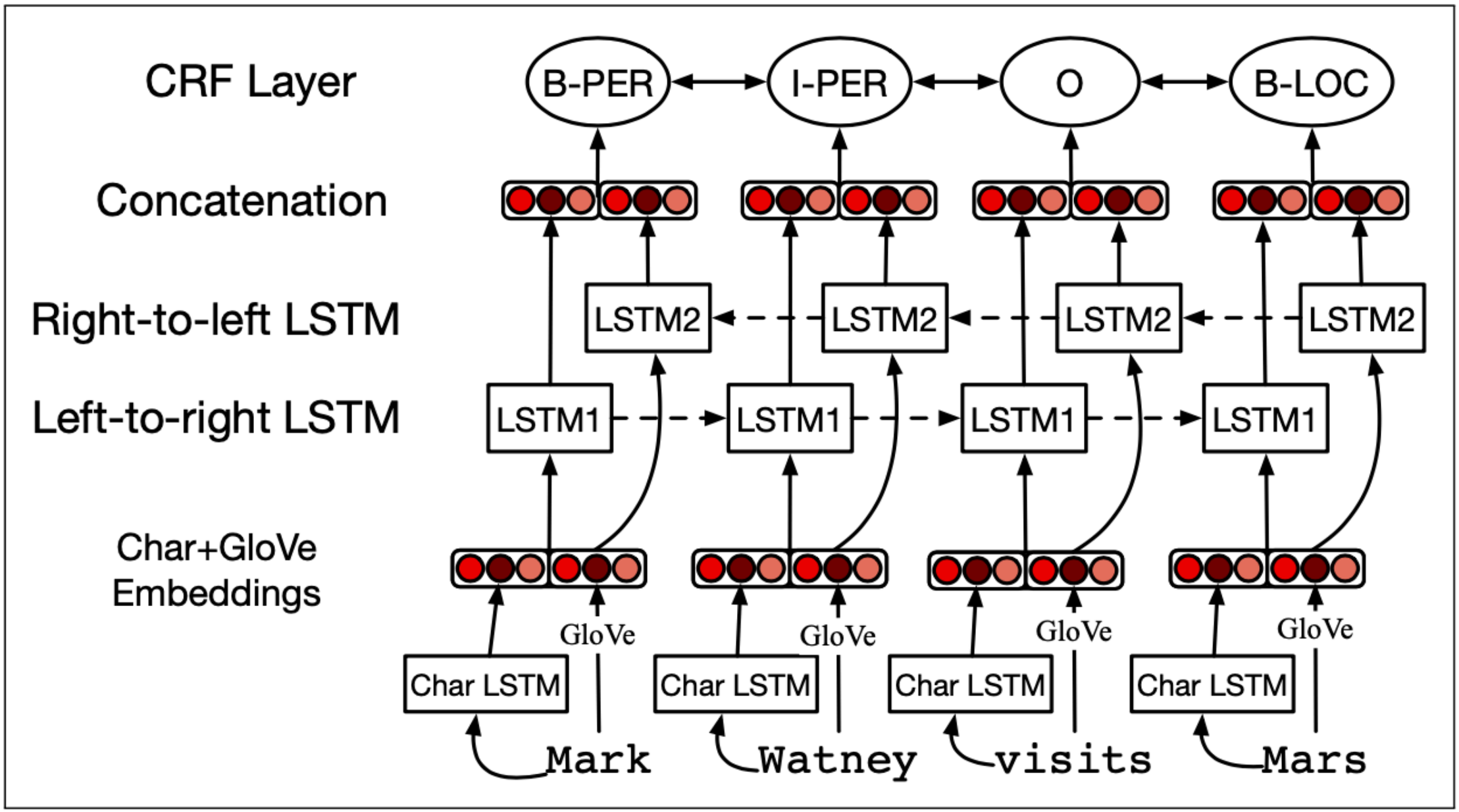


Figure 17.8 Putting it all together: character embeddings and words together a bi-LSTM sequence model. After (Lample et al., 2016)

Bidirectional LSTM models for NERC

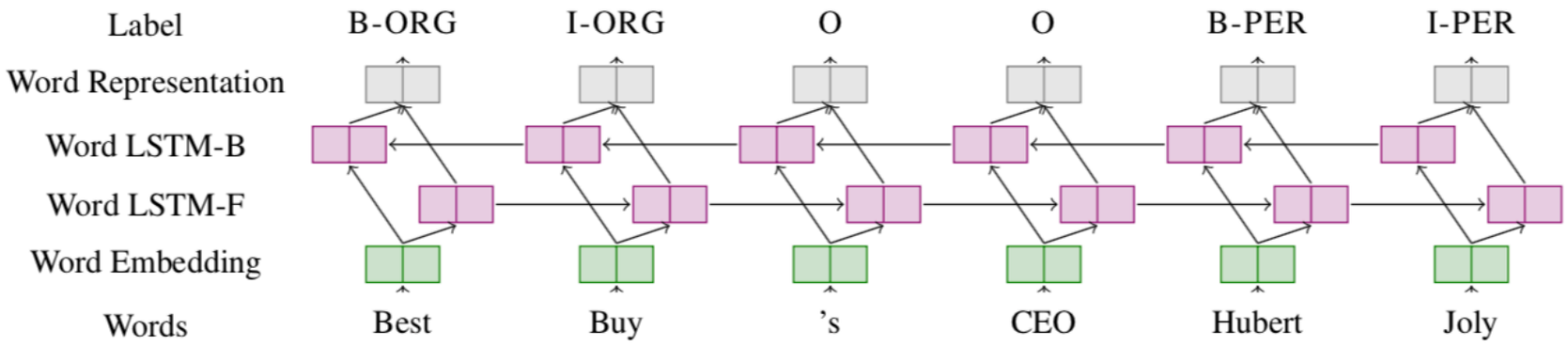


Figure 1: Word level NN architecture for NER

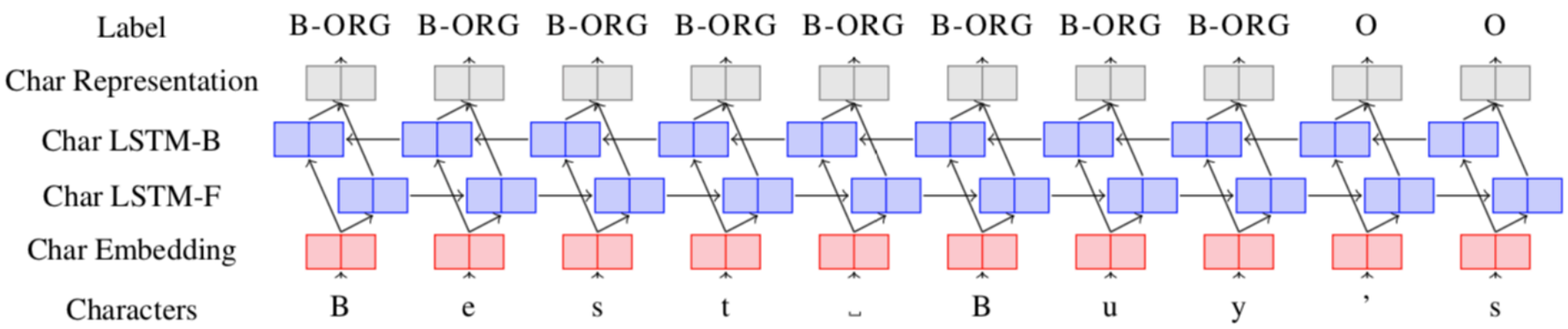


Figure 2: Character level NN architecture for NER

Yadav & Bethard 2018

https://towardsdatascience.com/besides-word-embedding-why-you-need-to-know-character-embedding-6096a34a3b10

Bidirectional LSTM models for NERC

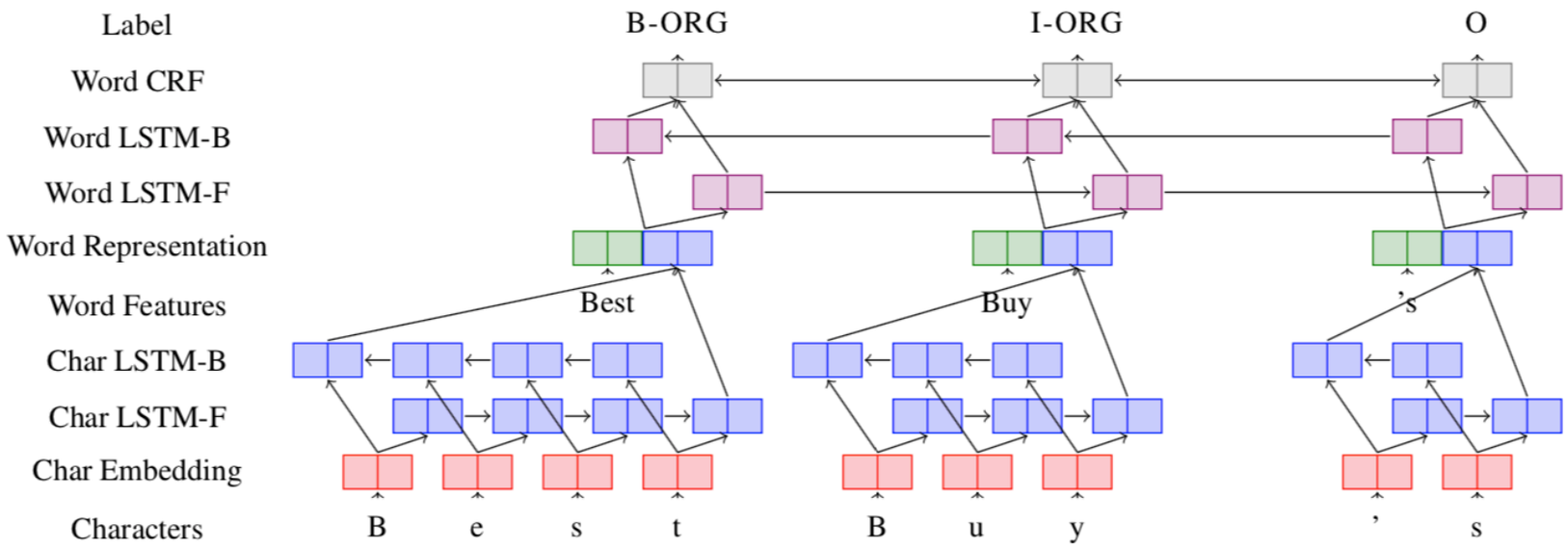


Figure 3: Word+character level NN architecture for NER

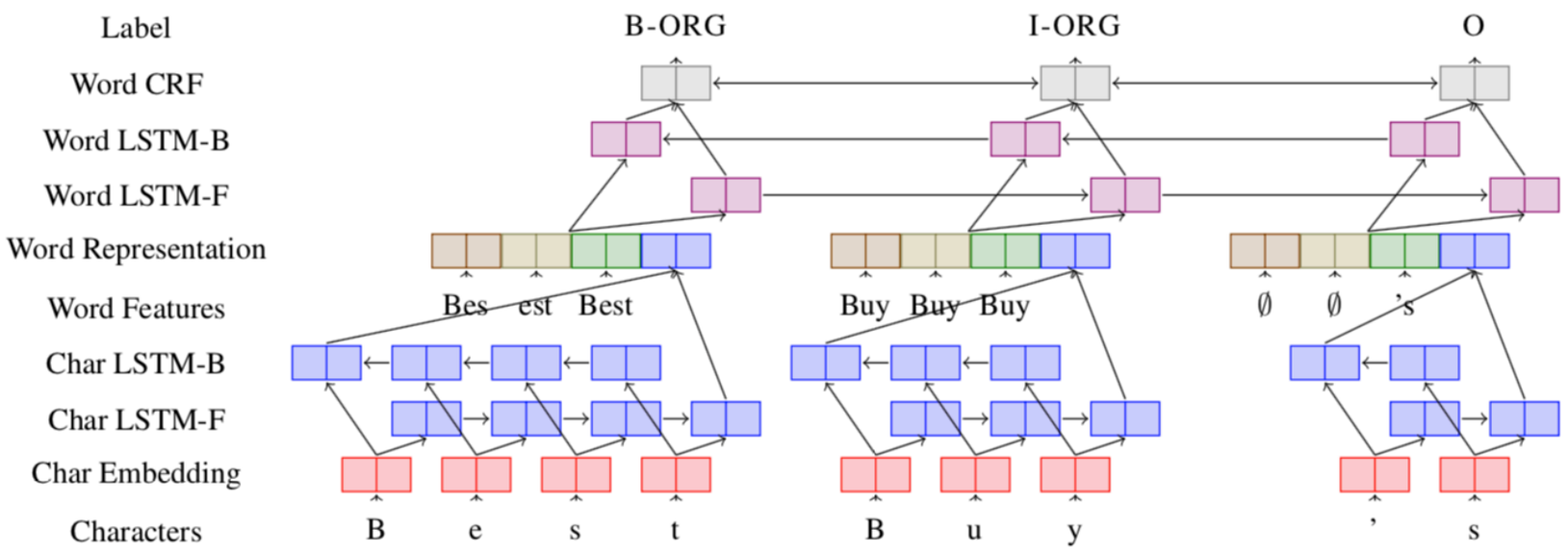


Figure 4: Word+character+affix level NN architecture for NER

Affix embeddings from all n-gram prefixes and suffixes of words in the training corpus

Yadav & Bethard 2018

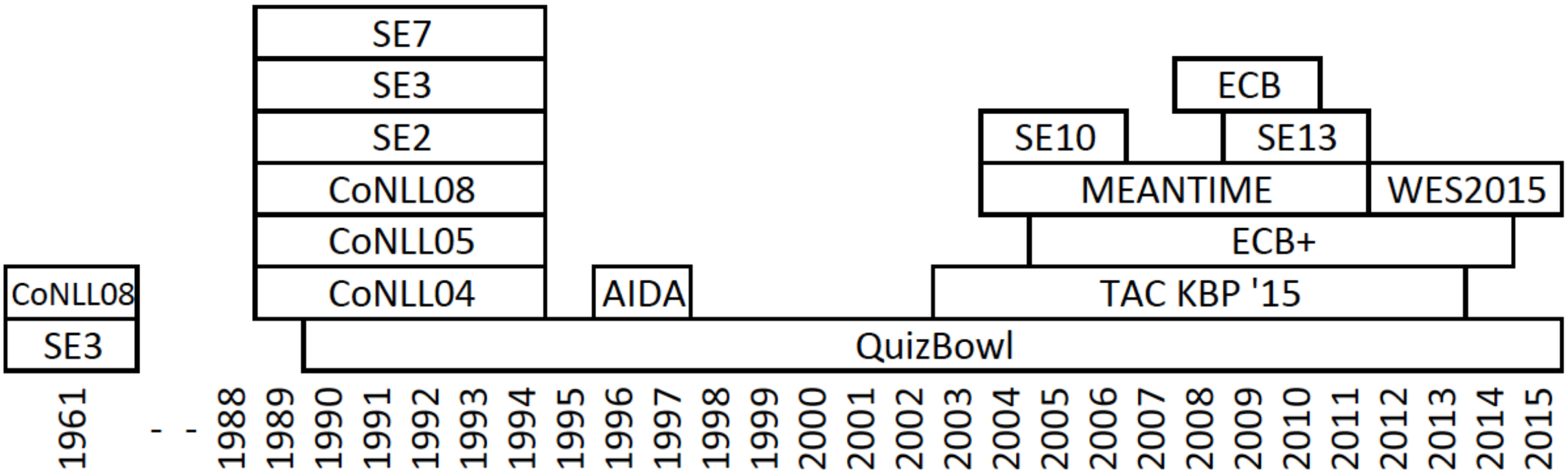
Factors that impact performance for NERC

- The annotation of the spans, annotation of nesting
  - [[[White House] [press] secretary] Scott McClellan]
  - [The [CEO] of the [US]-Based company [Facebook]]
- Type of text: news or tweets/ social media
- Entity types: people, organisations, amounts, dates, events
- Amount of training data
- Difference between training data and test data:
  - domain dependency of entities
  - training data rapidly becomes obsolete

Measuring performance for entities

- Is simple precision and recall enough?
- Neil Young & Crazy Horse
- Score per chunk (only give a true positive score if the entire NE is correctly classified)
    - Here “Neil” gets a false negative score, “&” gets a false positive score, “Horse” again a false negative.
- Some other metrics exist (e.g., MUC) that give partial credit (complex rules)

How specific is our data?



Ilievski, Postma, and Vossen, Semantic overfitting: what `world' do we consider when evaluating disambiguation of text? COLING 2016.

What is the effect of gazetteers on data over time?

Performance drops when shifting data

Agerri and Rigau 2016

Table 6: NERC CoNLL 2003 testb results.

	Precision	Recall	F1
Newsreader (ixa-pipe-nerc)	91.64	90.21	**90.92**
Stanford NER (CRF)	-	-	88.08
Ratinov et al. (2009)	-	-	90.57
Passos et al. (2014)	-	-	90.90

Table 7: NERC Intra-document Benchmarking with Wikinews.

System	mention extent	Precision	Recall	F1
Newsreader (ixa-pipe-nerc)	Inner phrase-based	62.15	76.06	**68.41**
Stanford NER (all english crf distsim)	Inner phrase-based	63.53	68.21	65.79
Newsreader (ixa-pipe-nerc)	Inner token-based	72.17	79.31	**75.57**
Stanford NER (all english crf distsim)	Inner token-based	77.14	71.77	74.36
Newsreader (ixa-pipe-nerc)	Outer phrase-based	53.01	68.03	**59.59**
Stanford NER (all english crf distsim)	Outer phrase-based	52.86	59.51	55.99
Newsreader (ixa-pipe-nerc)	Outer token-based	73.40	67.20	**70.16**
Stanford NER (all english crf distsim)	Outer token-based	78.22	60.63	68.31

P. Vossen, R. Agerri, I. Aldabe, A. Cybulska, M. van Erp, A. Fokkens, E. Laparra, A. Minard, A. P. Aprosio, G. Rigau, M. Rospocher, and R. Segers, “NewsReader: Using knowledge resources in a cross-lingual reading machine to generate more knowledge from massive streams of news”, Special issue knowledge-based systems, elsevier, 2016. dx.doi.org/10.1016/j.knosys.2016.07.013

Domain specific NER

	Dict	MedLine (80.10%)			DrugBank (19.90%)			Complete dataset		
		P	R	F1	P	R	F1	P	R	F1
Feature-engineered machine learning systems										
Rocktäschel et al. (2013)	Yes	60.70	55.80	58.10	88.10	87.50	87.80	73.40	69.80	71.50
Liu et al. (2015) (baseline)	No	-	-	-	-	-	-	78.41	67.78	72.71
Liu et al. (2015) (MED. emb.)	No	-	-	-	-	-	-	82.70	69.68	75.63
Liu et al. (2015) (state of the art)	Yes	78.77	60.21	68.25	90.60	88.82	**89.70**	84.75	72.89	**78.37**
NN word model										
Chalapathy et al. (2016) (relaxed performance)	No	52.93	52.57	52.75	87.07	83.39	85.19	-	-	-
NN word + character model Yadav et al. (2018)	No	73	62	67	87	86	87	79	72	75
NN word + character + affix model Yadav et al. (2018)	No	74	64	**69**	89	86	87	81	74	77

91+ on CoNLL 2003

90+ on CoNLL 2003

Table 2: DrugNER results on the MedLine and DrugBank test data (80.10% and 19.90% of the test data, respectively). The Yadav et al. (2018) experiments report no decimal places because they were run after the end of shared task, and the official evaluation script outputs no decimal places.

Yadav & Bethard 2018

Pragmatic approaches by companies

- First, use high-precision rules to tag unambiguous entity mentions.
- Then, search for substring matches of the previously detected names.
- Consult application-specific name lists to identify likely name entity mentions from the given domain.
- Finally, apply probabilistic sequence labeling techniques that make use of the tags from previous stages as additional features.

Juravsky & Martin, page 282

NERC References

- Yadav, Vikas, and Steven Bethard. "A survey on recent advances in named entity recognition from deep learning models." In Proceedings of the 27th International Conference on Computational Linguistics, pp. 2145-2158. 2018.
- Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami and Chris Dyer, 2016, Neural Architectures for Named Entity Recognition, NAACL.
- P. Vossen, R. Agerri, I. Aldabe, A. Cybulska, M. van Erp, A. Fokkens, E. Laparra, A. Minard, A. P. Aprosio, G. Rigau, M. Rospocher, and R. Segers, “NewsReader: Using knowledge resources in a cross-lingual reading machine to generate more knowledge from massive streams of news”, Special issue knowledge-based systems, elsevier, 2016. dx.doi.org/10.1016/j.knosys.2016.07.013
- Zhiheng Huang, Wei Xu, Kai Yu 2015, Bidirectional LSTM-CRF Models for Sequence Tagging, arXiv.1508.01991v1
- Ilievski, Postma, and Vossen, Semantic overfitting: what `world' do we consider when evaluating disambiguation of text? COLING 2016.
- Rodrigo Agerri and German Rigau. 2016. Robust multilingual named entity recognition with shallow semi- supervised features. Artificial Intelligence, 238:63–82.
- https://github.com/guillaumegenthial/tf_ner
- https://www.depends-on-the-definition.com/named-entity-recognition-conditional-random-fields-python/
- https://towardsdatascience.com/besides-word-embedding-why-you-need-to-know-character-embedding-6096a34a3b10
- https://machinelearningmastery.com/develop-character-based-neural-language-model-keras/
- https://www.kaggle.com/abhinavwalia95/entity-annotated-corpus

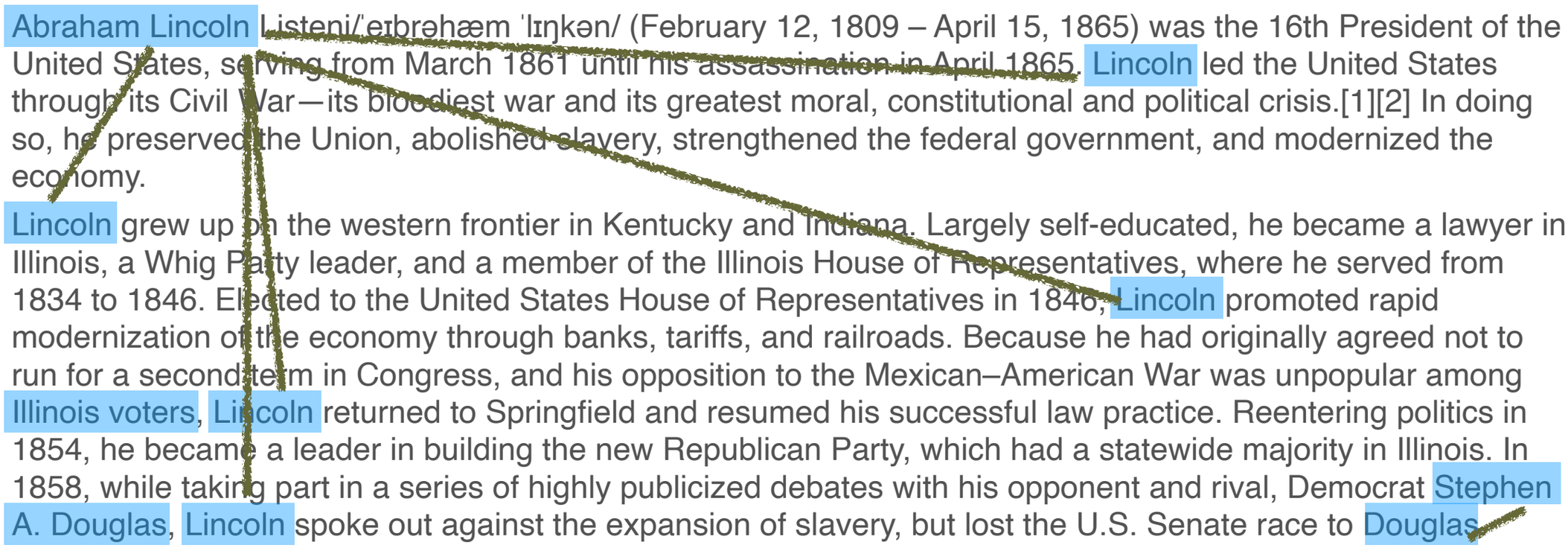
There is more than named entity expressions

- Identities: people with the same name (Joe Smith) are not necessarily people with the same identity

- Coreference: also phrases (“the president”) and pronouns (“he”, “she”) can make reference to the same entity

What is Coreference Resolution

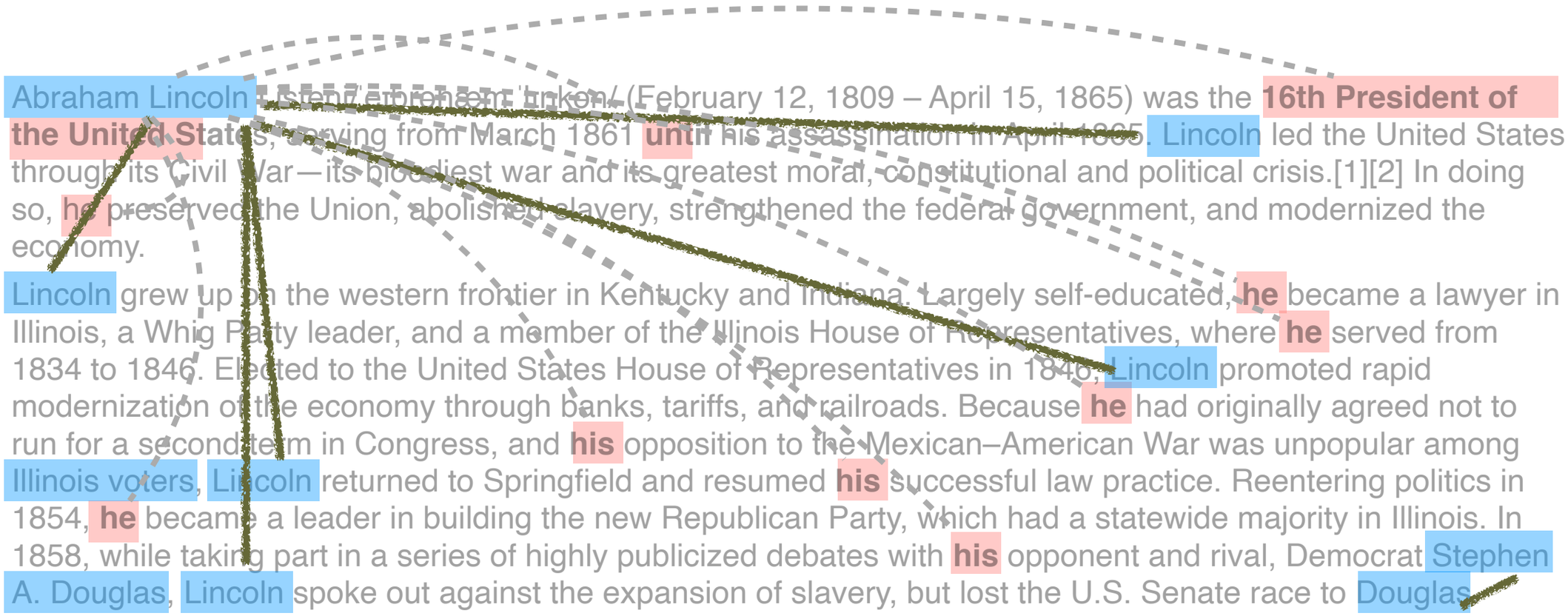
- Coreference resolution is the task of finding out which words/phrases refer to the same entity



source: http://en.wikipedia.org/wiki/Abraham_Lincoln

But it's actually more complicated

- Coreference resolution is the task of finding out which words/phrases refer to the same object



source: http://en.wikipedia.org/wiki/Abraham_Lincoln

But it’s actually more complicated

- Coreference resolution is the task of finding out which words/phrases refer to the same object

Abraham Lincoln Listeni/ˈeɪbrəhæm ˈlɪŋkən/ (February 12, 1809 – April 15, 1865) was the 16th President of the United States, serving from March 1861 until his assassination in April 1865. Lincoln led the United States through its Civil War—its bloodiest war and its greatest moral, constitutional and political crisis.[1][2] In doing so, he preserved the Union, abolished slavery, strengthened the federal government, and modernized the economy.

Lincoln grew up on the western frontier in Kentucky and Indiana. Largely self-educated, he became a lawyer in Illinois, a Whig Party leader, and a member of the Illinois House of Representatives, where he served from 1834 to 1846. Elected to the United States House of Representatives in 1846, Lincoln promoted rapid modernization of the economy through banks, tariffs, and railroads. Because he had originally agreed not to run for a second term in Congress, and his opposition to the Mexican–American War was unpopular among Illinois voters, Lincoln returned to Springfield and resumed his successful law practice. Reentering politics in 1854, he became a leader in building the new Republican Party, which had a statewide majority in Illinois. In 1858, while taking part in a series of highly publicized debates with his opponent and rival, Democrat Stephen A. Douglas, Lincoln spoke out against the expansion of slavery, but lost the U.S. Senate race to Douglas.

source: http://en.wikipedia.org/wiki/Abraham_Lincoln

How to do coreference resolution

Antecedent	Anaphor	Corefers?
Abraham Lincoln	He$_1$	yes
16th president of the United States	He$_1$	yes
Lincoln	His$_5$	yes
Stephen A. Douglas	He$_4$	no
Abraham Lincoln	Lincoln$_6$	yes
Member of the Illinois House of Representatives	He$_2$	yes

Lincoln grew up on the western frontier in Kentucky and Indiana. Largely self-educated, he$_1$ became a lawyer in Illinois, a Whig Party leader, and a member of the Illinois House of Representatives, where he$_2$ served from 1834 to 1846. Elected to the United States House of Representatives in 1846, Lincoln$_3$ promoted rapid modernization of the economy through banks, tariffs, and railroads. Because he$_4$ had originally agreed not to run for a second term in Congress, and his$_5$ opposition to the Mexican–American War was unpopular among Illinois voters, Lincoln$_6$ returned to Springfield and resumed his$_7$ successful law practice.

Why is coreference resolution important?

- Coreference is a frequently used natural language phenomenon
- Coreference resolution is essential to aggregate all knowledge and properties of the entities that a text makes reference to
- Coreference resolution is difficult because it stretches across sentences (syntax) and involves semantics and discourse

How to do coreference resolution (1)

Rule-based (Stanford multi-sieve, Lee eval 2013)

1. String matching

   - Will help you with proper names (*Smith* & *Smith*), common NPs (risky): *a man*, *another man*, *the man*
   - Partial matching is a problem (with/without titles?)
   - Fails on abbreviations and acronyms (and anything that doesn't use the same strings, e.g. *Lincoln, he*)

2. Agreement heuristics (anaphora must agree with their antecedents in name, gender and animacy)

3. Scoping (identify a text region where you expect to find the antecedent): Most recent matching subject is the most likely antecedent:

   - *John gave Bill a book. He ….*
   - *John gave Mary a book, She…*
   - *John gave Bill a book. Bill did not read it/ He did not read it/He asked him about the title*

How to do coreference resolution (2)

Machine Learning

- Annotated data marked up with co-reference chains
- Supervised technique to identify antecedents to anaphora
    - Clustering to merge pairwise coreference decisions into coreference chains
- Varied features used: part-of-speech tags, parse information, named entities, semantic class lookup, NP chunks, proximity, aliases, number, gender

Featured used for entity-coreference resolution

Type	Features
Mention	String match, part-of-speech, alias, number, gender (Soon, Ng, and Lim 2001), appositive, animacy, speaker (Lee et al. 2011), WordNet relation (Culotta, Wick, and McCallum 2007), modifier (Culotta, Wick, and McCallum 2007), overlap, quotation (Ng and Cardie 2002b), syntax subtree (Versley et al. 2008), dependency label (Björkelund and Nugues 2011), dependency path (Bergsma and Lin 2006), named-entity type (Denis and Baldridge 2009), semantic class (Soon, Ng, and Lim 2001), selectional preference (Dagan, Dagan, and Itai 1990), semantic roles (Ponzetto et al. 2006)
Textual context	Saliency (Lappin and Leass 1994), recency McCarthy (1996, pp. 87), narrative chain (Rahman and Ng 2012; Peng and Roth 2016)
Entity linking	Wikipedia (Ponzetto et al. 2006), Freebase attribute (Hajishirzi et al. 2013)

Table 1: A non-comprehensive list of features used in the literature. Each feature can be instantiated in many ways and sometimes one system contains more than one version.

Minh, Fokkens, Vossen, (under review), A Content-based Analysis of Coreference Resolution

Coreference performance

- CoNLL-2012 (Pradhan et al. 2012) standard benchmark in entity coreference resolution in recent years.:
    - 2,385 annotated English documents, totaling at 1.6M words, from various genres such as newswire, weblogs, and telephone conversations.
    - Highest reported result (after six years) is only 73% (Lee, He, and Zettlemoyer 2018).
- Some genres get much lower performance than others
    - Stanford Sieve (Lee et al. 2013) is lowest for newswire (55%) and highest for bible (67%).
    - The neural model of Clark and Manning (2016b) displays more than 10% difference between broadcast conversations (64%) and bible (78%).
- References:

    - Clark, Kevin and Christopher D. Manning. 2016a. Deep Reinforcement Learning for Mention-Ranking Coreference Models. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP-16), pages 2256–2262.
    - Clark, Kevin and Christopher D. Manning. 2016b. Improving Coreference Resolution by Learning Entity-Level Distributed Representations. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 643–653.
    - Lee, Heeyoung, Angel Chang, Yves Peirsman, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2013. Deterministic Coreference Resolution Based on Entity-Centric, Precision-Ranked Rules. Computational Linguistics, 39(4):885–916.
    - Lee, Kenton, Luheng He, and Luke Zettlemoyer. 2018. Higher-order Coreference Resolution with Coarse-to-fine Inference. pages 687–692.
    - Pradhan, Sameer, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 Shared Task: Modeling Multilingual Unrestricted Coreference in OntoNotes. EMNLP-CoNLL 2012.

Entity linking
What's in a name?

1. Task introduction: Who is Boris Johnson? Which Ford is Ford?
2. Phases of entity linking
3. Entity linkers
    a. Approaches
    b. Tools
4. Evaluation
    a. Aggregation
    b. Example

Entity tasks in NLP

- NER (Recognition): detecting the phrase that is the name of an entity

- NEC (Classification): assigning an entity type to the phrase

- NEL (Linking): establishing the identity of the entity in a given reference database (Wikipedia, DBpedia, YAGO)

- Coreference: any phrase that makes reference to an entity instance, including pronouns, noun phrases, abbreviations, acronyms, etc…

Task definition

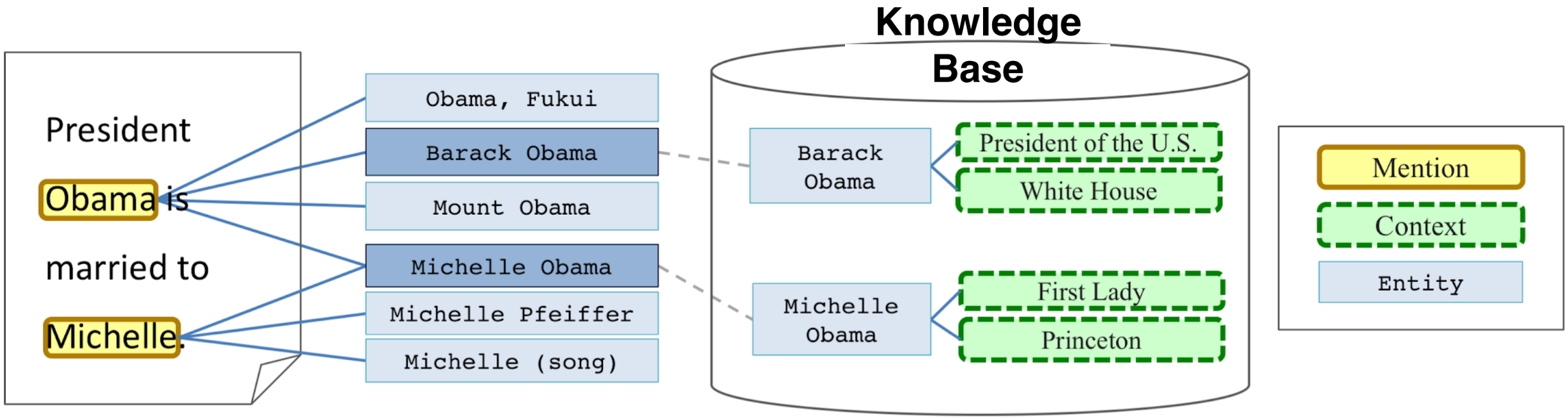
- Potentially ambiguous **entity mention** (“Paris”) needs to be linked to a canonical identifier/**instance** (http://dbpedia.org/resource/Paris) that fits the intended referent in the context of the text

- We find these instances in a **knowledge base**.

Example

“President **Obama** is married to **Michelle**.”

Example


President
Obama is
married to
Michelle.
Obama, Fukui
Barack Obama
Mount Obama
Michelle Obama
Michelle Pfeiffer
Michelle (song)
Knowledge Base
Barack Obama
President of the U.S.
White House
Michelle Obama
First Lady
Princeton
Mention
Context
Entity

Knowledge base

A catalog of things, usually entities. Each one has:

- **a unique identifier or record number, possibly a Unique Resource Identifier (URI):**
  - https://www.wikidata.org/wiki/Q513, https://en.wikipedia.org/wiki/Mount_Everest
- **one or more names**
  - **Mount Everest** -> “Mount Everest”, “Everest”, “Mount Qomolangma”, “Mt. Qomolangma”, “Mount Sagarmatha”, “Qomolangma”, “Chomolangma”, “Mt. Everest”, ...
- **other attributes**
  - Elevation: 8,848m
  - Coordinates: 27°59'17"N, 86°55'31"E
- **Connections to other entities**
  - Continent: Asia
  - Country: China, Country: Nepal
- **Textual description**
  - **example**

Usually a knowledge base has some of these aspects, but not all.

Knowledge base types

The knowledge bases can be classified into two types:

- Unstructured (e.g., Wikipedia)
    - Mostly contains a textual (“unstructured”) description
- Structured (e.g., Wikidata, DBpedia, ...)
    - Contains structured description of an entity
    - Property-value pairs

Unstructured knowledge bases: Wikipedia

https://en.wikipedia.org/wiki/Mount_Everest

Mount Everest

From Wikipedia, the free encyclopedia

Coordinates: 27°59′17″N 86°55′31″E

"Everest" redirects here. For other uses, see Everest (disambiguation).

This article's **tone or style may not reflect the encyclopedic tone used on Wikipedia**. See Wikipedia's guide to writing better articles for suggestions. *(October 2017) (Learn how and when to remove this template message)*

Mount Everest, known in Nepali as **Sagarmatha** (सगरमाथा) and in Tibetan as **Chomolungma** (ཇོ་མོ་གླང་མ), is Earth's highest mountain above sea level, located in the Mahalangur Himal sub-range of the Himalayas. The international border between Nepal (Province No. 1) and China (Tibet Autonomous Region) runs across its summit point.

The current official elevation of 8,848 m (29,029 ft), recognized by China and Nepal, was established by a 1955 Indian survey and subsequently confirmed by a Chinese survey in 1975.[1] In 2005, China remeasured the rock height of the mountain, with a result of 8844.43 m (29,017 ft). There followed an argument between China and Nepal as to whether the official height should be the rock height (8,844 m., China) or the snow height (8,848 m., Nepal). In 2010, an agreement was reached by both sides that the height of Everest is 8,848 m, and Nepal recognizes China's claim that the rock height of Everest is 8,844 m.[5]

In 1865, Everest was given its official English name by the Royal Geographical Society, upon a recommendation by Andrew Waugh, the British Surveyor General of India. As there appeared to be several different local names, Waugh chose to name the mountain after his predecessor in the post, Sir George Everest, despite Everest's objections.[6]

Mount Everest attracts many climbers, some of them highly experienced mountaineers. There are two main climbing routes, one approaching the summit from the southeast in Nepal (known as the "standard route") and the other from the north in Tibet. While not posing substantial technical climbing challenges on the standard route, Everest presents dangers such as altitude sickness, weather, and wind, as well as significant hazards from avalanches and the Khumbu Icefall. As of 2017, nearly 300 people have died on Everest, many of whose bodies remain on the mountain.[7]

The first recorded efforts to reach Everest's summit were made by British mountaineers. As Nepal did not allow foreigners into the country at the time, the British made several attempts on the north ridge route from the Tibetan side. After the first reconnaissance expedition by the British in 1921 reached 7,000 m (22,970 ft) on the North Col, the 1922 expedition pushed the north ridge route up to 8,320 m (27,300 ft), marking the first time a human had climbed above 8,000 m (26,247 ft). Seven porters were killed in an avalanche on the descent from the North Col. The 1924 expedition resulted in one of the greatest mysteries on Everest to this day: George Mallory and Andrew Irvine made a final summit attempt on 8 June but never returned, sparking debate as to whether or not they were the first to reach the top. They had been spotted high on the mountain that day but disappeared in the clouds, never to be seen again, until Mallory's body was found in 1999 at 8,155 m (26,755 ft) on the north face. Tenzing Norgay and Edmund Hillary made the first official ascent of Everest in 1953, using the southeast ridge route. Norgay had reached 8,595 m (28,199 ft) the previous year as a member of the 1952 Swiss expedition. The Chinese mountaineering team of Wang Fuzhou, Gonpo, and Qu Yinhua made the first reported ascent of the peak from the north ridge on 25 May 1960.[8][9]

Contents [hide]

Mount Everest

Mount Everest as viewed from Kalapatthar.

Highest point	
Elevation	8,848 metres (29,029 ft)[1] Ranked 1st
Prominence	Ranked 1st (Notice special definition for Everest)
Listing	Seven Summits Eight-thousander Country high point Ultra
Coordinates	27°59′17″N 86°55′31″E[2]
Naming	

Structured knowledge bases: DBpedia & Wikidata

dbpedia.org/page/Mount_Everest

DBpedia | Browse using | Formats

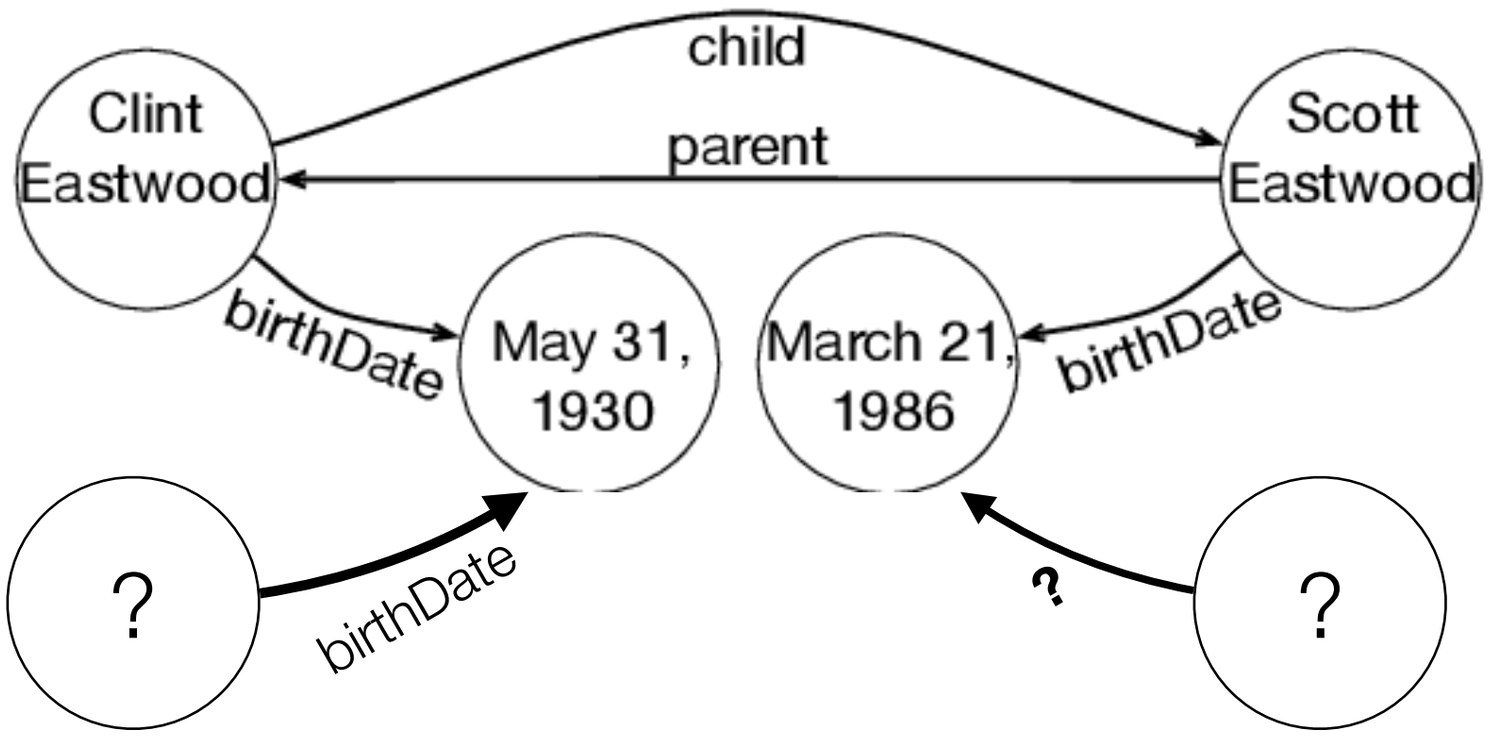
geo:geometry	POINT(86.925277709961 27.988056182861)
geo:lat	27.988056 (xsd:float)
geo:long	86.925278 (xsd:float)
prov:wasDerivedFrom	wikipedia-en:Mount_Everest?oldid=744845387
foaf:depiction	wiki-commons:Special:FilePath/Mount-Everest.jpg
foaf:isPrimaryTopicOf	wikipedia-en:Mount_Everest
foaf:name	Mount Everest (en)
is dbo:deathPlace of	dbr:Shailendra_Kumar_Upadhyaya dbr:Mick_Burke_(mountaineer) dbr:Ray_Genet dbr:Scott_Fischer dbr:Karl_Gordon_Henize dbr:Hristo_Prodanov dbr:David_Sharp_(mountaineer) dbr:Rob_Hall dbr:Pasang_Lhamu_Sherpa dbr:Zygmunt_Andrzej_Heinrich dbr:Mohammad_Khaled_Hossain dbr:Andrew_Irvine_(mountaineer) dbr:Maurice_Wilson

https://www.wikidata.org/wiki/Q513

instance of	mountain 0 references
part of	Seven Summits 1 reference Himalayas 1 reference
image	 Mount Everest from Rongbuk may 2005.JPG 3,008 × 2,000; 1.12 MB 0 references

Structured KBs are essentially graph networks

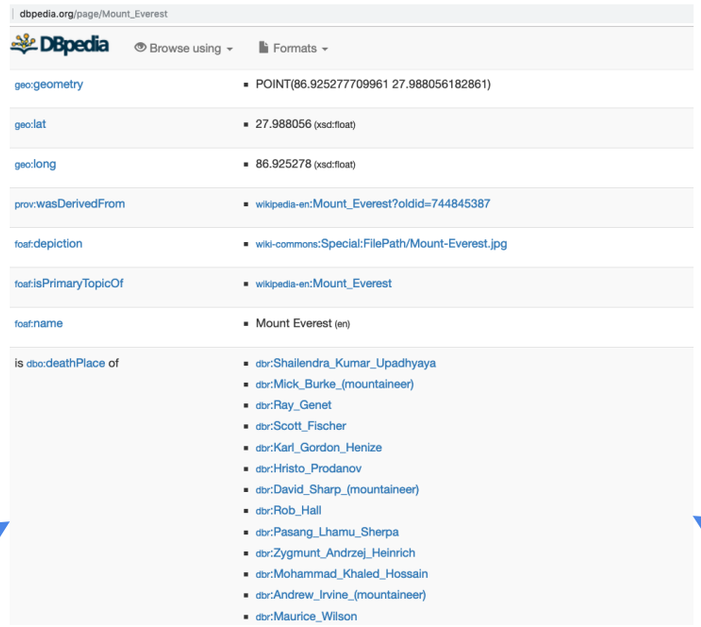
... with billions (!) of such facts

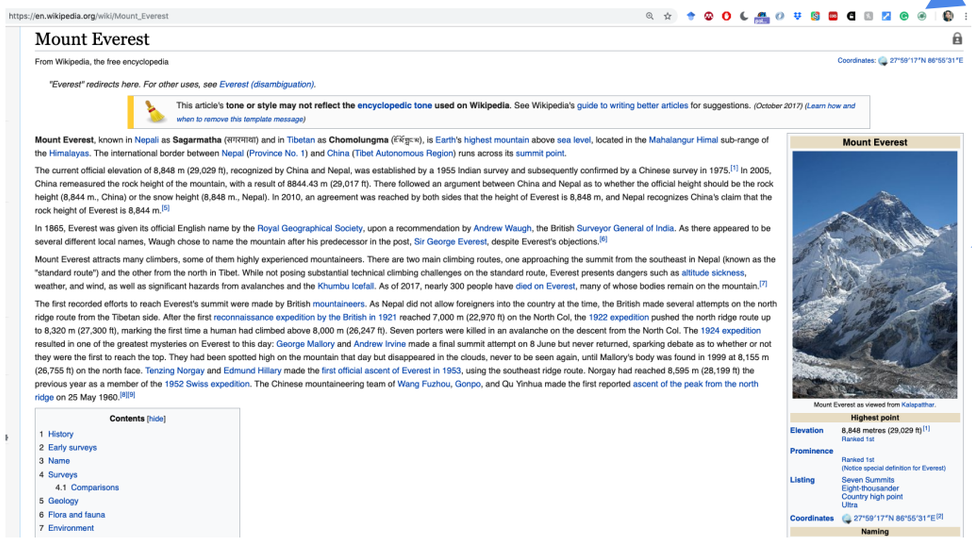


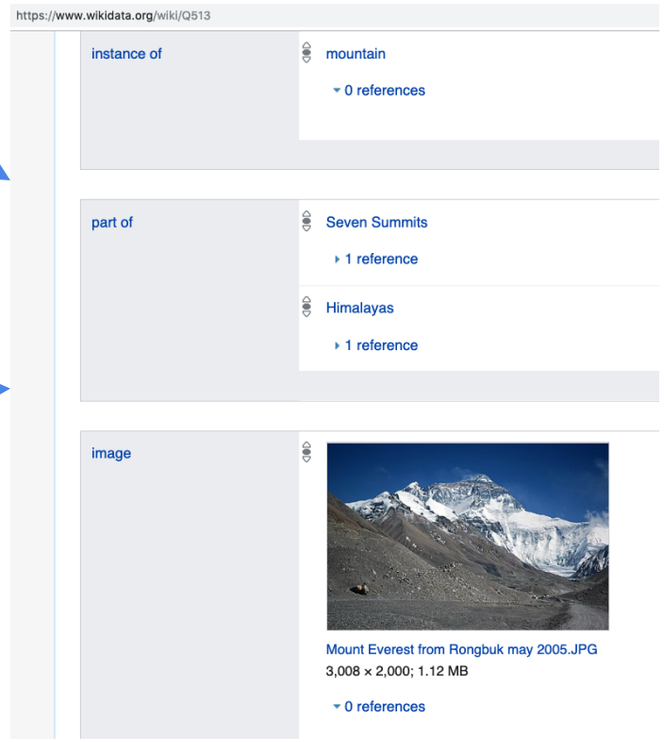
Who else has the same birth date?

What else is linked to March, 21, 1986?

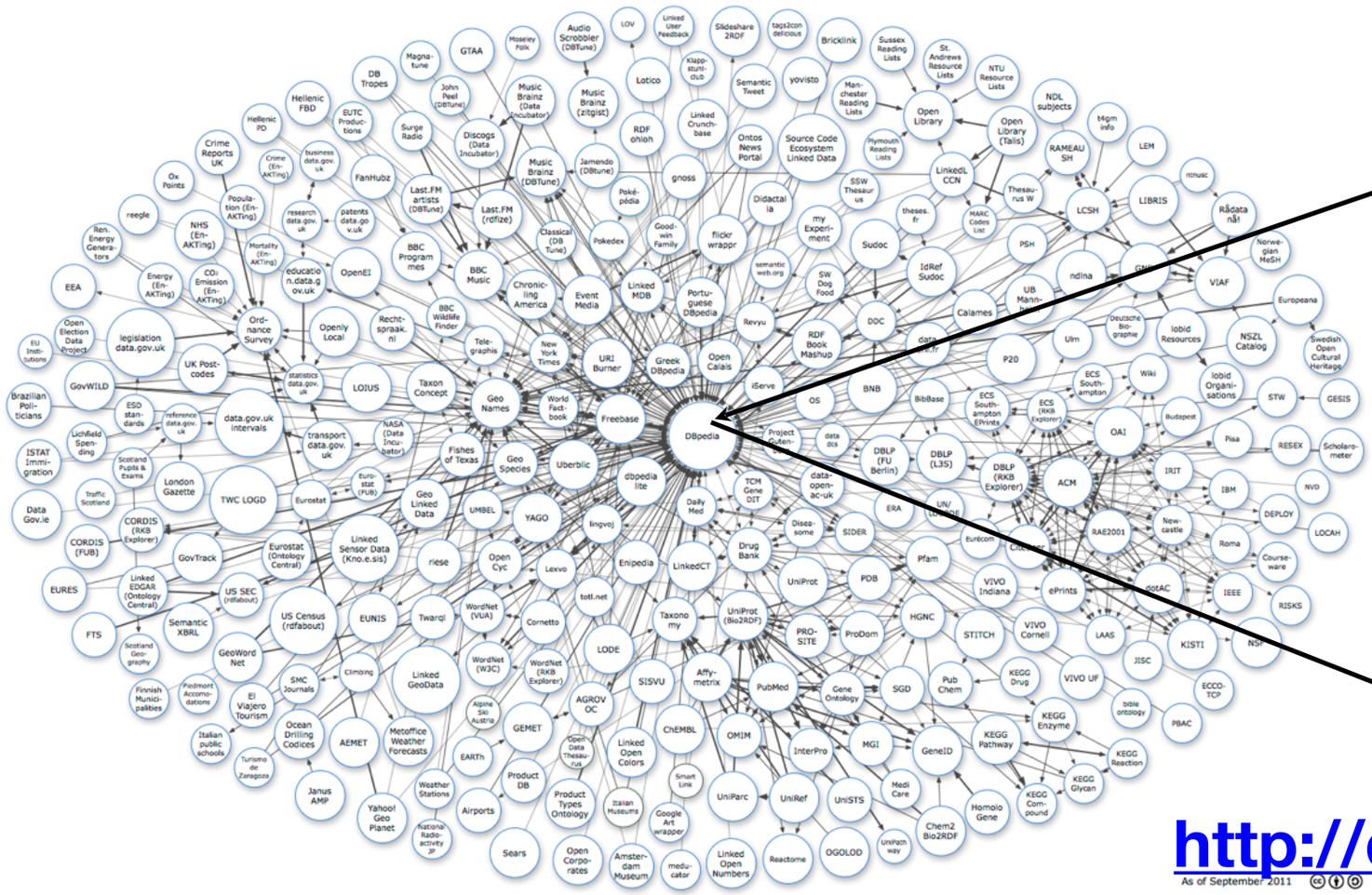
Knowledge bases are also connected to each other







Many more KBs exist, and they are connected -> the LOD Cloud



“John Travolta”

http://dbpedia.org/resource/John_Travolta

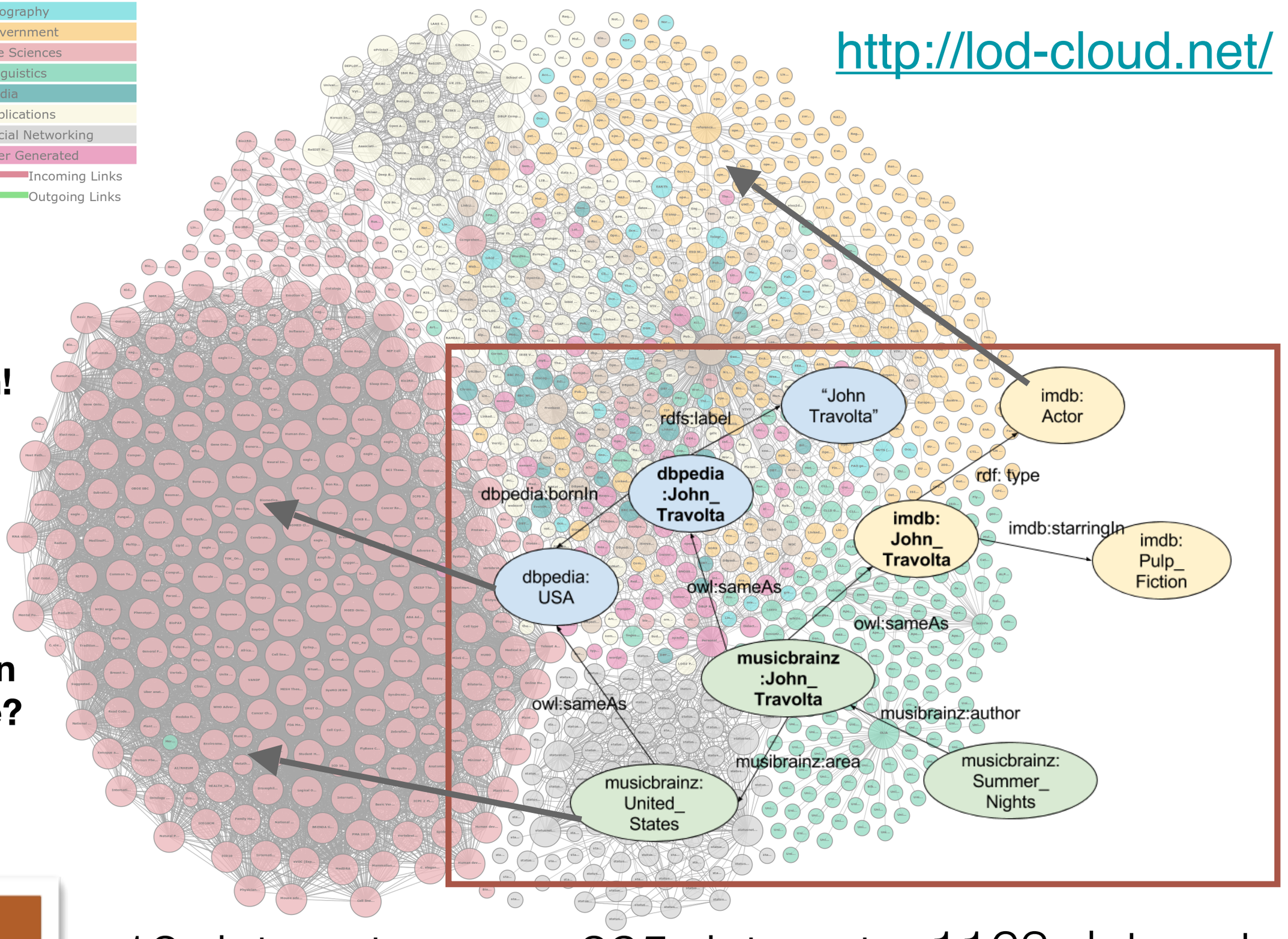
Legend

- Cross Domain
- Geography
- Government
- Life Sciences
- Linguistics
- Media
- Publications
- Social Networking
- User Generated
- Incoming Links
- Outgoing Links

http://lod-cloud.net/

I love John!

Which John do you love?



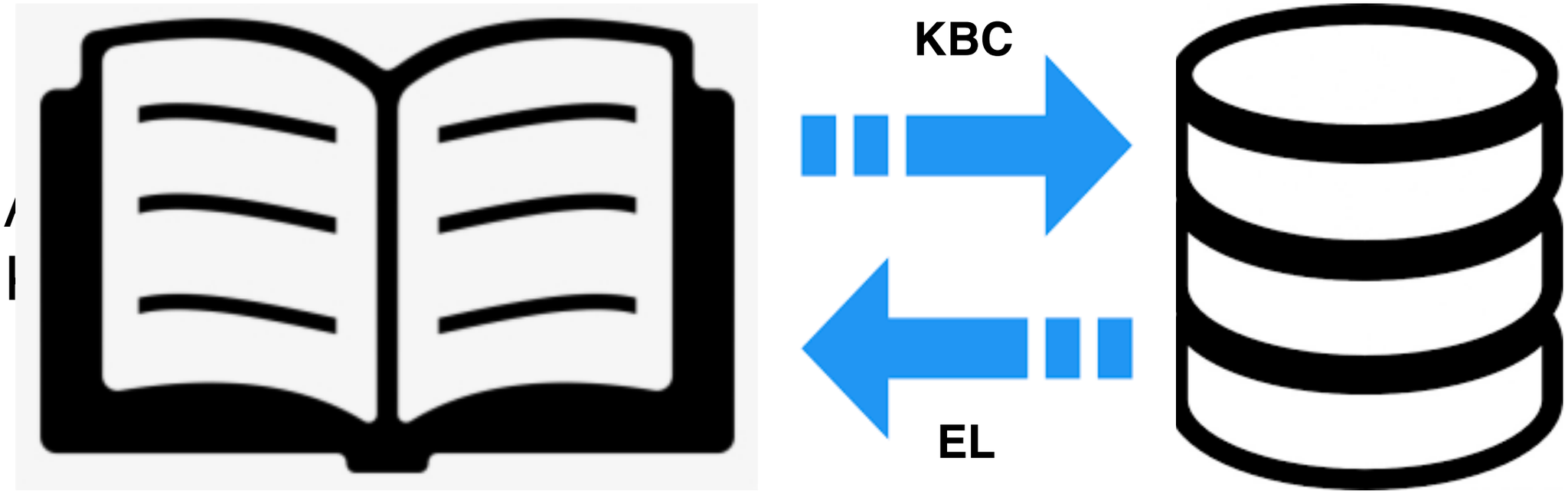
Power of the division of labour

12 data sets
2 billion triples
2007

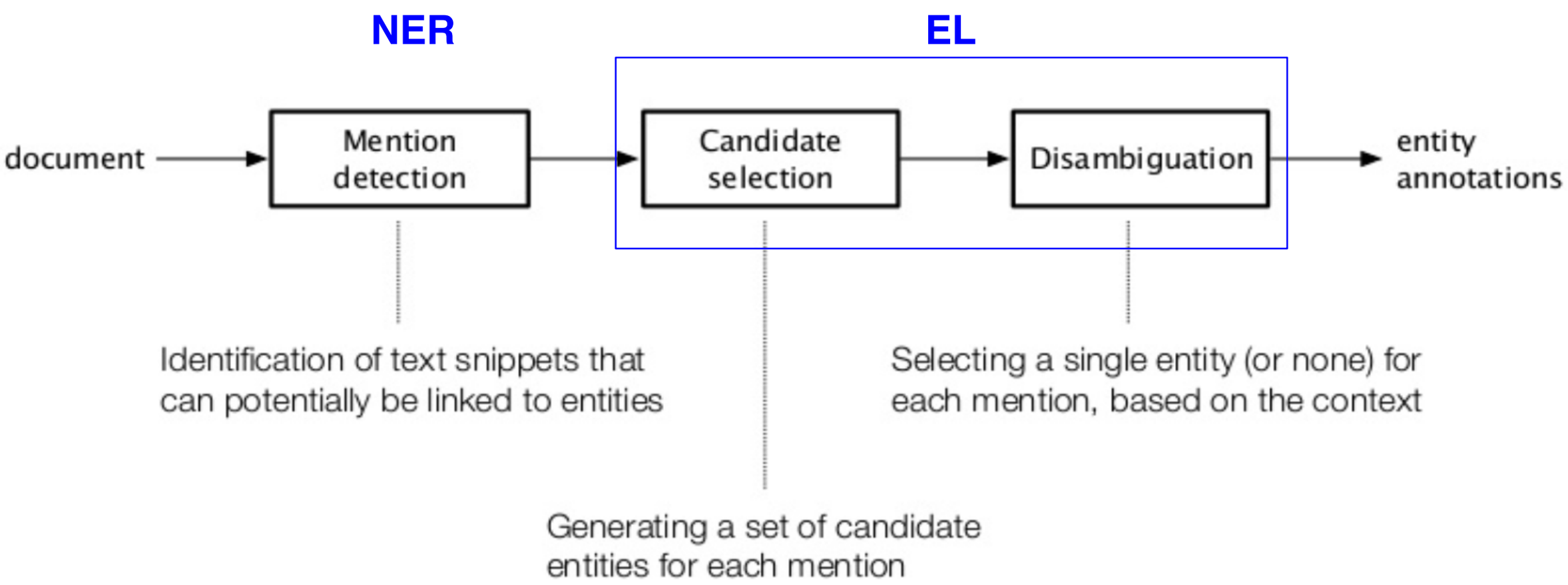
295 data sets
31 billion triples
2011

1163 data sets
2017

Other benefits of connecting text and knowledge bases



Anatomy of an entity linking system



Source: https://www.slideshare.net/krisztianbalog/entity-linking-65308055

Candidate generation & selection

- For each of the recognised mentions in text, get the potential referents (instances) in a knowledge base (KB), following the “closed world assumption” (=the world is in the KB).
- The goal is to balance between generating too many candidates (too much ‘noise’) and generating too little candidates (missing the correct one)
- Trade-off between precision and recall
- Candidate generation is an art by itself!

But... how do you choose the top X (or 30) candidates?

- We need a way to rank them somehow.
- A common ranking criteria is **commonness**: for a given mention in Wikipedia texts, how often (relatively) does it refer to some instance in Wikipedia.
- For example, of all the mentions of “Germany” in Wikipedia, what is the percentage that refers to the country vs the football club vs the handball club vs the government vs etc.
    - Perform the ranking of candidate entities based on their overall popularity, i.e., "most common sense"

$$P(e|m) = \frac{n(m,e)}{\sum_{e'} n(m,e')}$$

$n(m,e)$ → the number of times entity e is the link destination of mention m

$\sum_{e'} n(m,e')$ → total number of times mention m appears as a link

Example

Entity	Commonness
FIFA_World_Cup	0.2358
FIS_Apline_Ski_World_Cup	0.0682
2009_FINA_Swimming_World_Cup	0.0633
World_Cup_(men's_golf)	0.0622
...	

Bulgaria's best **World Cup** performance was in the **1994 World Cup** where they beat **Germany**, to reach the semi-finals, losing to Italy, and finishing in fourth ...

Entity	Commonness
1998_FIFA_World_Cup	0.9556
1998_IAAF_World_Cup	0.0296
1998_Alpine_Skiing_World_Cup	0.0059
...	

Entity	Commonness
Germany	0.9417
Germany_national_football_team	0.0139
Nazi_Germany	0.0081
German_Empire	0.0065
...	

Also, observe:

- Dominance within a form
- Topical bias

In practice, about 30 candidates per mention is enough.

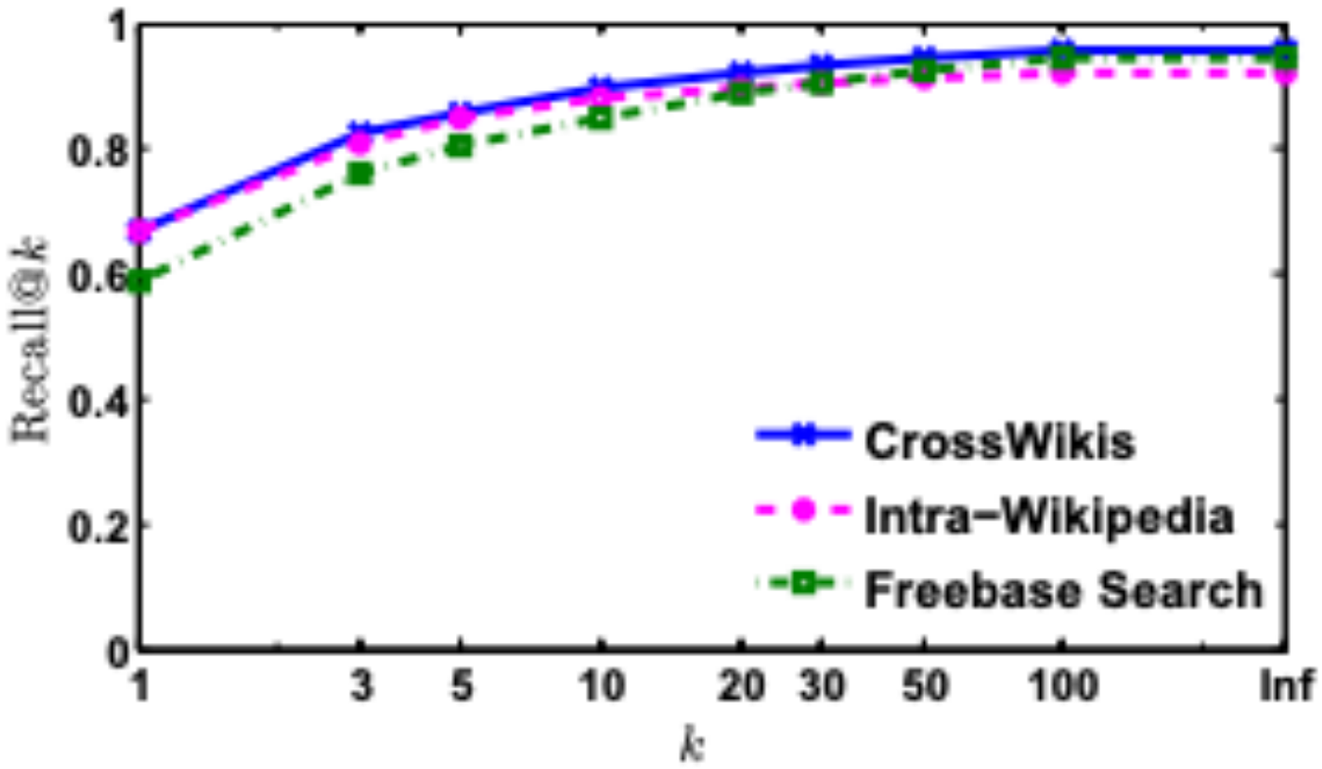
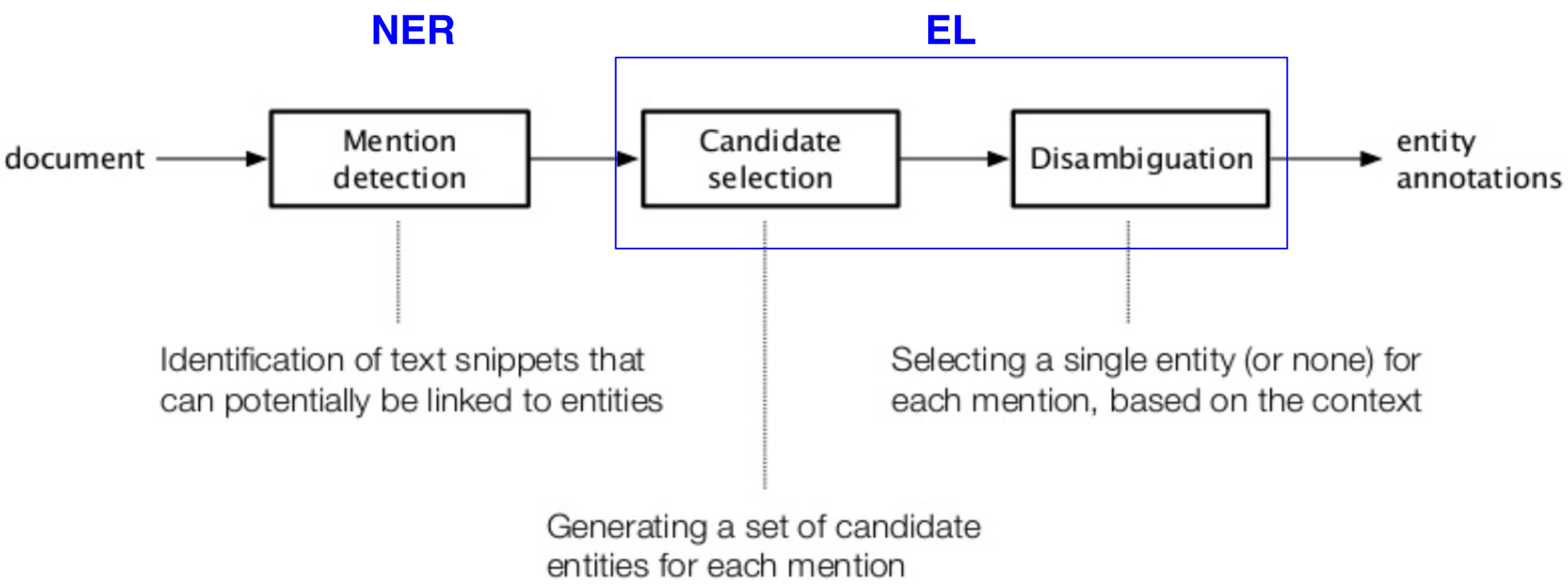


Figure 3: Recall@k on an aggregate of nine data sets, comparing three **candidate generation** methods.

- CrossWikis=counts of web anchors from Google crawl (175M entries)
- IntraWikipedia=counts of wikipedia anchor links
- Freebase=Google knowledge base, returns 220 candidates per query

Source: https://www.mitpressjournals.org/doi/pdfplus/10.1162/tacl_a_00141

Anatomy of an entity linking system



Source: https://www.slideshare.net/krisztianbalog/entity-linking-65308055

Disambiguation or linking

Goal: decide which of the candidates (or none) is the correct referent.

When would this phase be easy and when difficult?

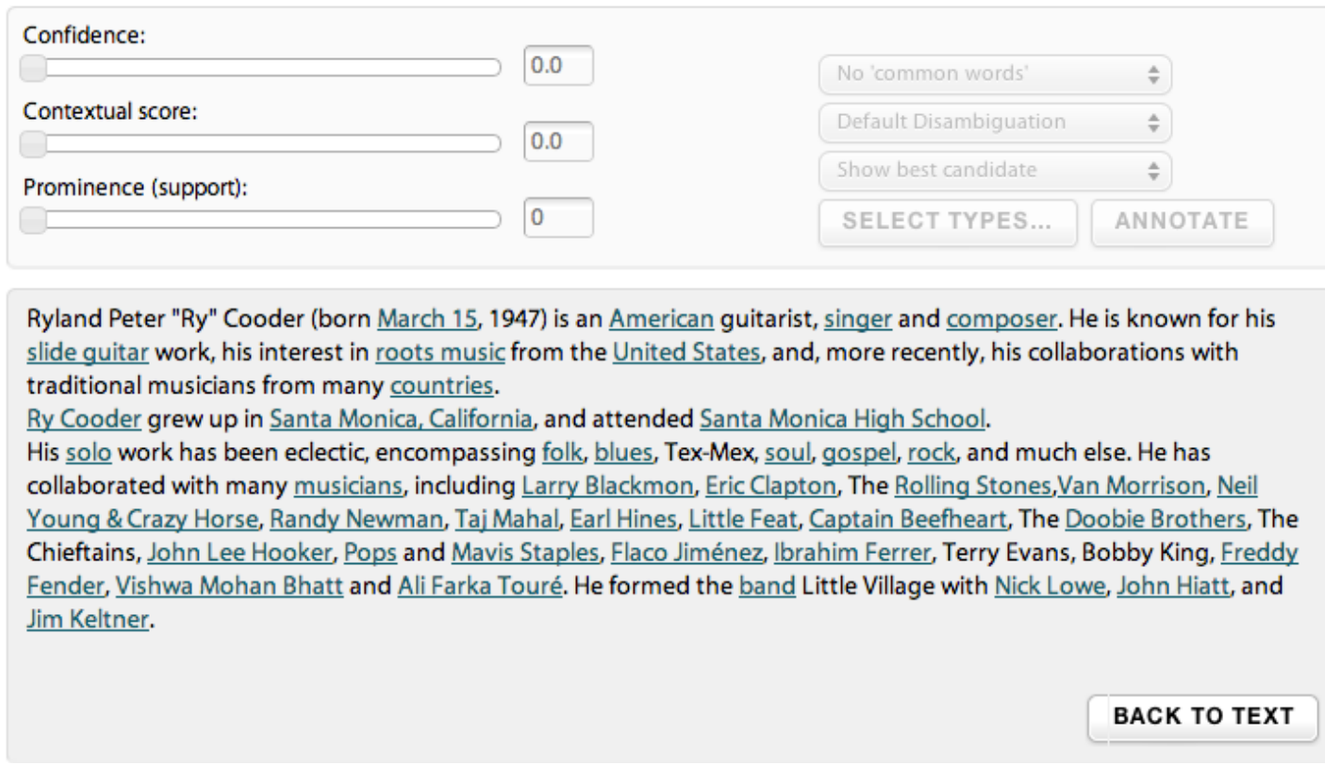
Entity linking methods

	Word-based	**Graph-based**
Main idea	Find the candidate with the most similar description to the one of a mention in text	Find candidates that are coherent with each other according to connections in the KB
Scoring example	measure text similarity, combine with **TF/IDF** weighting to measure relevance of a word	Put all candidates with their facts in a graph network and prune until only one candidate per mention is left
Decision unit	individual/local	collective/global
KB	unstructured (Wikipedia)	Structured (DBpedia, etc.)
Example	DBpedia Spotlight, Wikifier	AIDA/AGDISTIS

TF/IDF = term frequency * inverse document frequency
Measures the degree to which terms are important for a document, based on the frequency in the document but normalised by checking if it occurs in all documents or just a few

3a. Word-based methods: DBpedia Spotlight

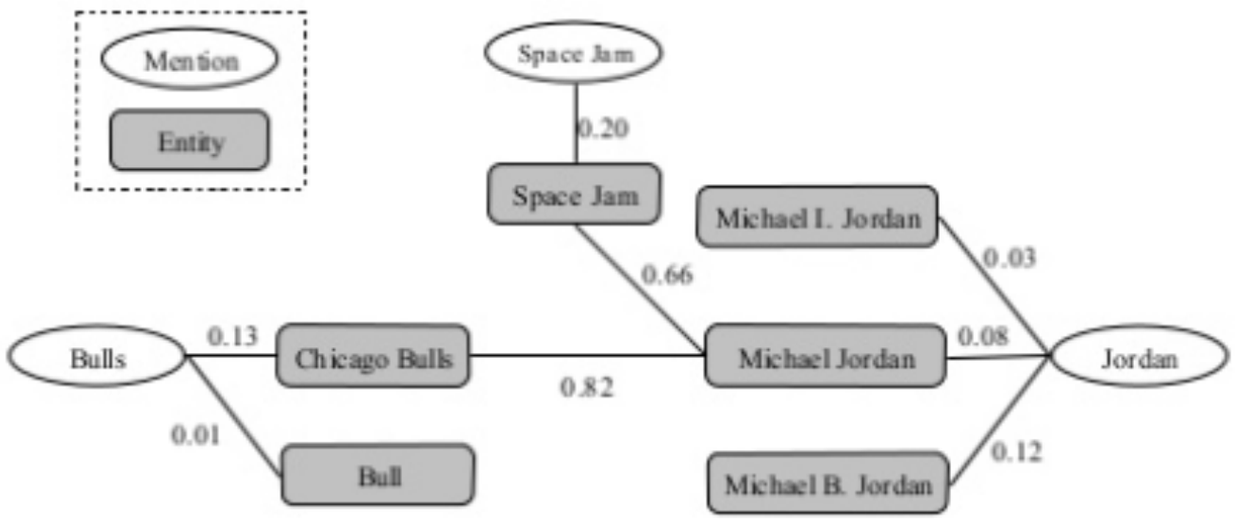
- Compute cosine similarity between the text paragraph with an entity mention and Wikipedia descriptions of each candidate.
- Decide for one mention at a time.
- The linking can be restricted to certain types or even to a custom set of entities.



3b. Graph-based methods: AIDA and AGDISTIS

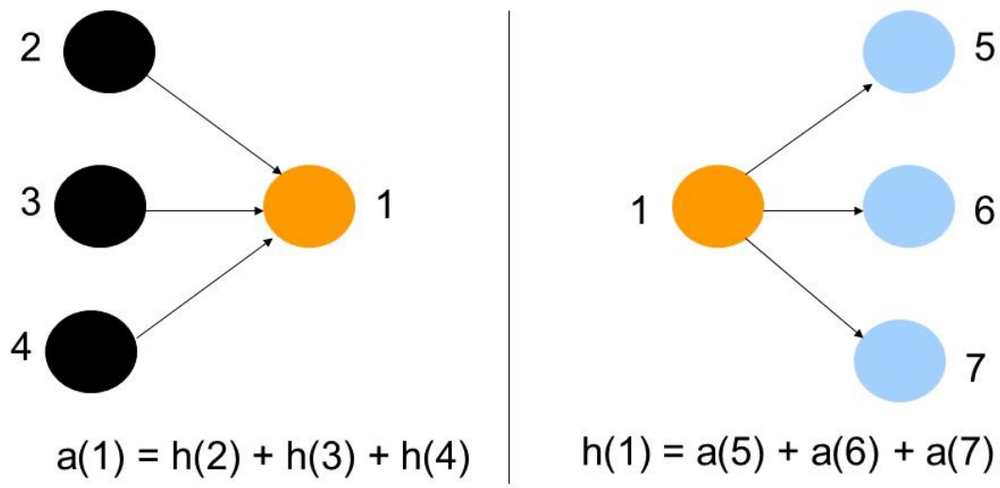
1. Construct a subgraph that contains all entity candidates with some facts from a KB.
2. Find the best connected candidates per mention:

Example



Compute relatedness between the candidates (AIDA)

Authority and Hubness



Find the “hubs” in the graph (AGDISTIS)

Local vs global disambiguation

- Note that the idea in the graph-based approaches is to make the optimal global decision (we disambiguate all entities together).

- This is different than in DBpedia Spotlight, where we disambiguate entities one by one.

Datasets for Entity Linking

Group	Data Set	# of Mentions	Entity Types	KB	# of NILs	Eval. Metric
UIUC	ACE	244	Any Wikipedia Topic	Wikipedia	0	BOC F1
	MSNBC	654	Any Wikipedia Topic	Wikipedia	0	BOC F1
AIDA	AIDA-dev	5917	PER,ORG,LOC,MISC	Yago	1126	Accuracy
	AIDA-test	5616	PER,ORG,LOC,MISC	Yago	1131	Accuracy
TAC KBP	TAC09	3904	PER^T,ORG^T,GPE	TAC $\subset$ Wiki	2229	Accuracy
	TAC10	2250	PER^T,ORG^T,GPE	TAC $\subset$ Wiki	1230	Accuracy
	TAC10T	1500	PER^T,ORG^T,GPE	TAC $\subset$ Wiki	426	Accuracy
	TAC11	2250	PER^T,ORG^T,GPE	TAC $\subset$ Wiki	1126	B^3+ F1
	TAC12	2226	PER^T,ORG^T,GPE	TAC $\subset$ Wiki	1049	B^3+ F1

Table 1: Characteristics of the nine NEL data sets. Entity types: The AIDA data sets include named entities in four NER classes, Person (PER), Organization (ORG), Location (LOC) and Misc. In TAC KBP data sets, both Person (PER^T) and Organization entities (ORG^T) are defined differently from their NER counterparts and geo-political entities (GPE), different from LOC, exclude places like `KB:Central California`. KB (Sec. 2.2): The knowledge base used when each data was being developed. Evaluation Metric (Sec. 2.3): Bag-of-Concept F1 is used as the evaluation metric in (Ratinov et al., 2011; Cheng and Roth, 2013). B^3+ F1 used in TAC KBP measures the accuracy in terms of entity clusters, grouped by the mentions linked to the same entity.

Vinculum: ablation study

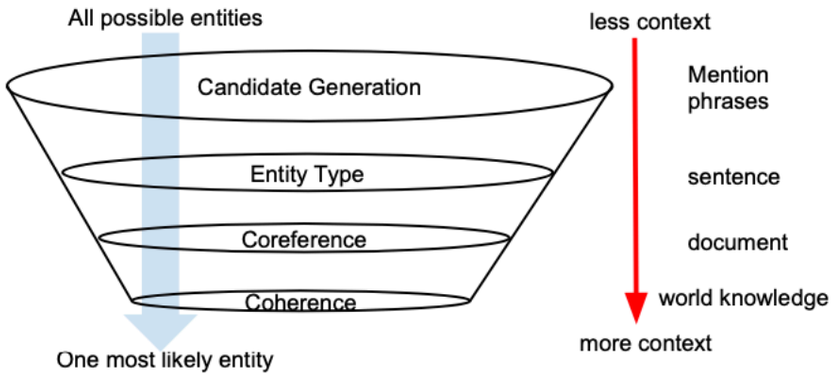


Figure 2: The process of finding the best entity for a mention. All possible entities are sifted through as VINCULUM proceeds at each stage with a widening range of context in consideration.

	ACE		MSNBC		AIDA-dev		AIDA-test	
	R	P	R	P	R	P	R	P
NER	89.7	10.9	77.7	65.5	89.0	75.6	87.1	74.0
+NP	96.0	2.4	90.2	12.4	94.7	21.2	92.2	21.8
+DP	96.8	1.8	90.8	9.3	95.8	14.0	93.8	13.5
+NP+DP	98.0	1.2	92.0	5.8	95.9	9.4	94.1	9.4

Table 3: Performance(%, R: Recall; P: Precision) of the correct mentions using different **mention extraction** strategies. ACE and MSNBC only annotate a subset of all the mentions and therefore the absolute values of precision are largely underestimated.

mention detection

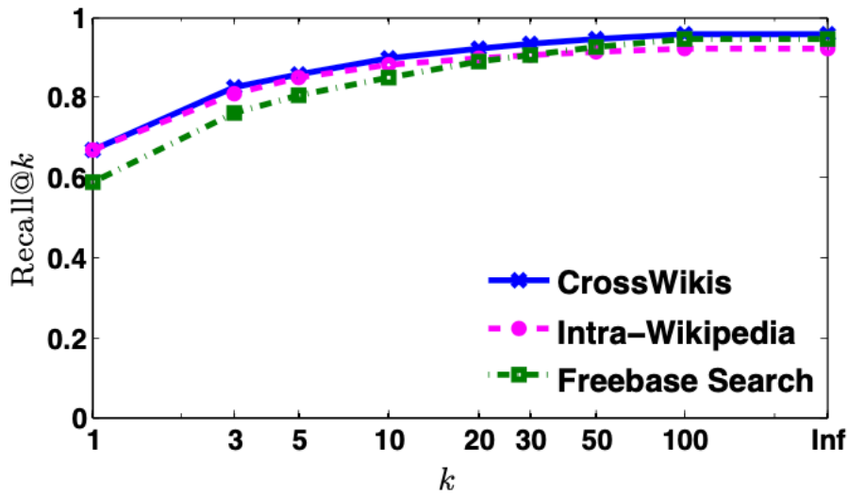


Figure 3: Recall@k on an aggregate of nine data sets, comparing three **candidate generation** methods.

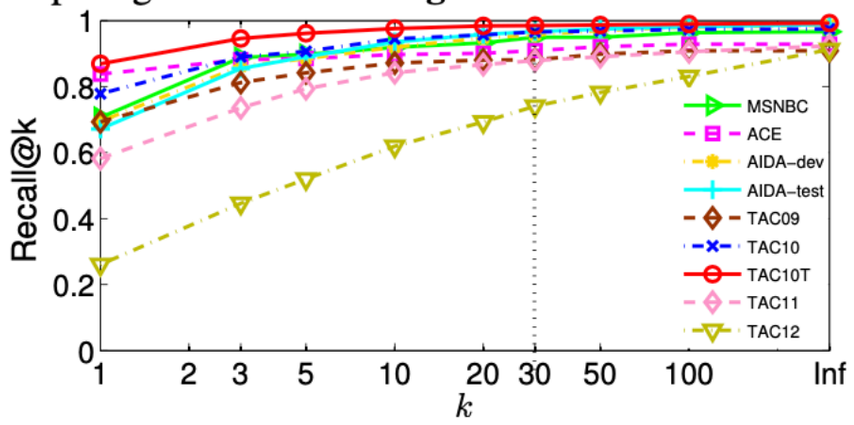


Figure 4: Recall@k using CrossWikis for candidate generation, split by data set. 30 is chosen to be the cut-off value in consideration of both efficiency and accuracy.

candidate generation

System performance

baseline

types

Approach	TAC09	TAC10	TAC10T	TAC11	TAC12	AIDA-dev	AIDA-test	ACE	MSNBC
CrossWikis only	80.4	85.6	86.9	78.5	62.4	62.6	60.4	87.7	70.3
+NER	79.2	83.3	85.1	76.6	61.1	66.4	**66.2**	77.0	71.8
+FIGER	**81.0**	**86.1**	**86.9**	**78.8**	**63.5**	**66.7**	64.6	**87.7**	**75.4**
+NER(GOLD)	**85.7**	87.4	88.0	80.1	**66.7**	72.6	72.0	89.3	83.3
+FIGER(GOLD)	84.1	**88.8**	**89.0**	**81.6**	66.1	**76.2**	**76.5**	**91.8**	**87.4**

Table 4: Performance (%) after **incorporating entity types**, comparing two sets of entity types (NER and FIGER). Using a set of fine-grained entity types (FIGER) generally achieves better results.

coherence

Approach	TAC09	TAC10	TAC10T	TAC11	TAC12	AIDA-dev	AIDA-test	ACE	MSNBC
no COH	80.9	86.2	87.0	78.6	59.9	68.9	66.3	87.7	86.6
+NGD	**81.8**	85.7	86.8	79.7	63.2	**69.5**	**67.7**	88.1	86.8
+REL	81.2	86.3	87.0	79.3	63.1	69.1	66.4	**88.5**	86.1
+BOTH	81.4	**86.8**	**87.0**	**79.9**	**63.7**	69.4	67.5	**88.5**	**86.9**

Table 5: Performance (%) after re-ranking candidates using coherence scores, comparing two **coherence measures** (NGD and REL). "no COH": no coherence based re-ranking is used. "+BOTH": an average of two scores is used for re-ranking. Coherence in general helps: a combination of both measures often achieves the best effect and NGD has a slight advantage over REL.

Error analysis

Error Category	TAC09	TAC10	TAC10T	TAC11	TAC12	AIDA-dev	AIDA-test	ACE	MSNBC
Metonymy	16.7%	0.0%	3.3%	0.0%	0.0%	60.0%	60.0%	5.3%	20.0%
Wrong Entity Types	13.3%	23.3%	20.0%	6.7%	10.0%	6.7%	10.0%	31.6%	5.0%
Coreference	30.0%	6.7%	20.0%	6.7%	3.3%	0.0%	0.0%	0.0%	20.0%
Context	30.0%	26.7%	26.7%	70.0%	70.0%	13.3%	16.7%	15.8%	15.0%
Specific Labels	6.7%	36.7%	16.7%	10.0%	3.3%	3.3%	3.3%	36.9%	25.0%
Misc	3.3%	6.7%	13.3%	6.7%	13.3%	16.7%	10.0%	10.5%	15.0%
# of examined errors	30	30	30	30	30	30	30	19	20

Table 9: **Error analysis:** We analyze a random sample of 250 of VINCULUM's errors, categorize the errors into six classes, and display the frequencies of each type across the nine datasets.

Category	Example	Gold Label	Prediction
Metonymy	South Africa managed to avoid a fifth successive defeat in 1996 at the hands of the All Blacks ...	South Africa national rugby union team	South Africa
Wrong Entity Types	Instead of Los Angeles International, for example, consider flying into Burbank or John Wayne Airport ...	Bob Hope Airport	Burbank, California
Coreference	It is about his mysterious father, Barack Hussein Obama, an imperious if alluring voice gone distant and then missing.	Barack Obama Sr.	Barack Obama
Context	Scott Walker removed himself from the race, but Green never really stirred the passions of former Walker supporters, nor did he garner outsized support "outstate".	Scott Walker (politician)	Scott Walker (singer)
Specific Labels	What we like would be Seles , (Olympic champion Lindsay) Davenport and Mary Joe Fernandez .	1996 Summer Olympics	Olympic Games
Misc	NEW YORK 1996-12-07	New York City	New York

Table 8: We divide linking errors into **six error categories** and provide an example for each class.

Entity linking pipelines

	VINCULUM	AIDA	WIKIFIER
Mention Extraction	NER	NER	NER, noun phrases
Candidate Generation	**CrossWikis**	an intra-Wikipedia dictionary	an intra-Wikipedia dictionary
Entity Types	**FIGER**	NER	NER
Coreference	find the representative mention	-	re-rank the candidates
Coherence	link-based similarity, relation triples	link-based similarity	link-based similarity, relation triples
Learning	unsupervised	trained on AIDA	trained on a Wikipedia sample

Table 7: Comparison of entity linking pipeline architectures. VINCULUM components are described in detail in Section 4, and correspond to Figure 2. Components found to be most useful for VINCULUM are highlighted.

Approach	TAC09	TAC10	TAC10T	TAC11	TAC12	AIDA-dev	AIDA-test	ACE	MSNBC	Overall
CrossWikis	80.4	85.6	86.9	78.5	62.4	62.6	62.4	87.7	70.3	75.0
+FIGER	81.0	86.1	86.9	78.8	63.5	66.7	64.5	87.7	75.4	76.7
+Coref	80.9	86.2	87.0	78.6	59.9	68.9	66.3	87.7	86.6	78.0
+Coherence =VINCULUM	**81.4**	**86.8**	**87.0**	79.9	63.7	69.4	67.5	**88.5**	86.9	79.0
AIDA	73.2	78.6	77.5	68.4	52.0	71.9	**74.8**	77.8	75.4	72.2
WIKIFIER	79.7	86.2	86.3	**82.4**	**64.7**	**72.1**	69.8	85.1	**90.1**	**79.6**

Table 6: **End-to-end performance** (%): We compare VINCULUM in different stages with two state-of-the-art systems, AIDA and WIKIFIER. The column "Overall" lists the average performance of nine data sets for each approach. CrossWikis appears to be a strong baseline. VINCULUM is 0.6% shy from WIKIFIER, each winning in four data sets; AIDA tops both VINCULUM and WIKIFIER on AIDA-test.

Further reading

- *Mendes, P. N., Jakob, M., García-Silva, A., & Bizer, C. (2011, September). DBpedia spotlight: shedding light on the web of documents. In Proceedings of the 7th international conference on semantic systems (pp. 1-8). ACM.*

- *Hoffart, J., Yosef, M. A., Bordino, I., Fürstenau, H., Pinkal, M., Spaniol, M., ... & Weikum, G. (2011, July). Robust disambiguation of named entities in text. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (pp. 782-792). Association for Computational Linguistics.*

- *Ling, X., Singh, S., & Weld, D. S. (2015). Design challenges for entity linking. Transactions of the Association for Computational Linguistics, 3, 315-328.*

- *Van Erp, M., Mendes, P., Paulheim, H., Ilievski, F., Plu, J., Rizzo, G., Waitelonis, J. (2016). Evaluating Entity Linking: An Analysis of Current Benchmark Datasets and a Roadmap for Doing a Better Job. LREC*