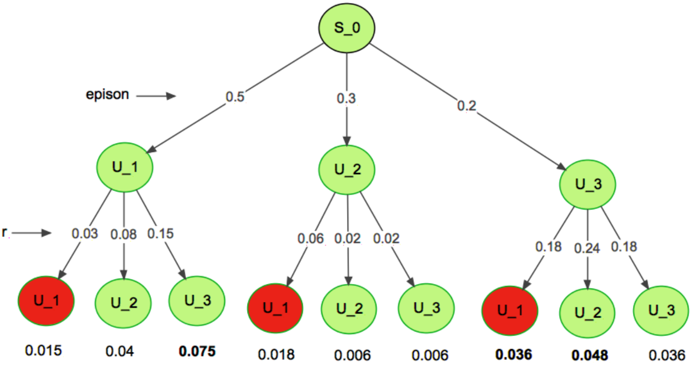
Text Mining

March 4 2019

Viterbi Algorithm for the Urn problem (first two symbols)



- At every stage, we only keep three nodes
- at the end of observation sequence - we have three nodes (total nodes - 3 x 8)
- complexity comes down from $|S|^{|o|}$ to $|S|.|o|$

HMM - POS Tagging

Goal: choose the most probable tag sequence given the observation sequence of n words $\hat{w}_1^n$

$$\hat{t}_1^n = \underset{t_1^n}{\operatorname{argmax}} P(t_1^n | w_1^n)$$

Using Bayes' rule

$$\hat{t}_1^n = \underset{t_1^n}{\operatorname{argmax}} \frac{P(w_1^n | t_1^n) P(t_1^n)}{P(w_1^n)}$$

Simplifying further by dropping the denominator

$$\hat{t}_1^n = \underset{t_1^n}{\operatorname{argmax}} P(w_1^n | t_1^n) P(t_1^n)$$

HMM - POS Tagging

HMM makes two further assumptions:

1. probability of a word depends only on its tag and is independent of neighbouring words and tags

$$P(w_1^n|t_1^n) \approx \prod_{i=1}^{n} P(w_i|t_i)$$

2. probability of a word depends only on its tag and is independent of neighbouring words and tags

$$P(t_1^n) \approx \prod_{i=1}^{n} P(t_i|t_{i-1})$$

Using these simplifications:

$$\hat{t}_1^n = \underset{t_1^n}{\operatorname{argmax}} P(t_1^n|w_1^n) \approx \underset{t_1^n}{\operatorname{argmax}} \prod_{i=1}^{n} \overbrace{P(w_i|t_i)}^{\text{emission}} \overbrace{P(t_i|t_{i-1})}^{\text{transition}}$$

HMM - POS Tagging

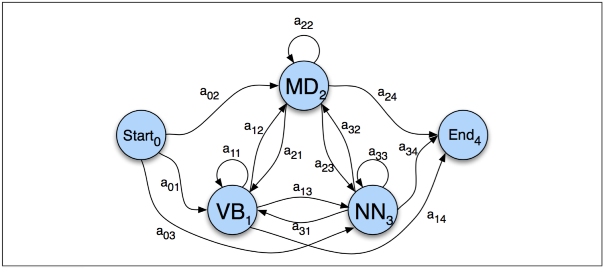


Figure: Markov chain corresponding to the hidden states of HMM. The transition probabilities A are used to compute the prior probability.

HMM - POS Tagging

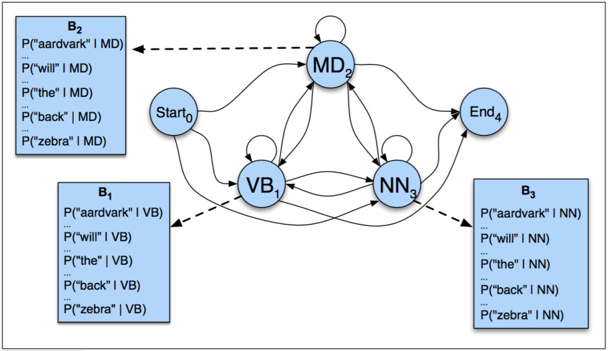


Figure: Observation likelihoods B for the HMM.

HMM - POS Tagging

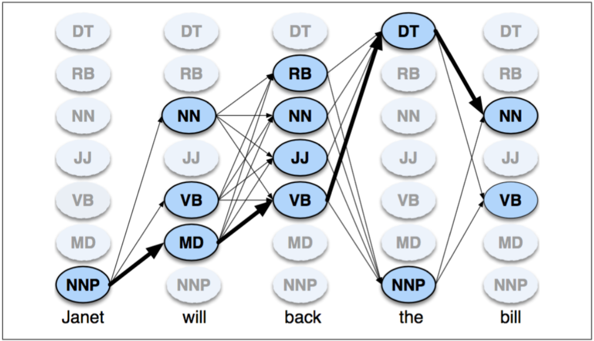
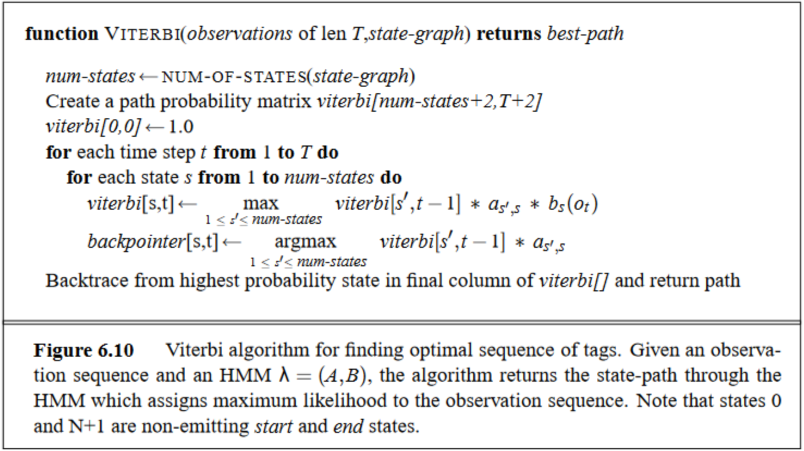


Figure: Observation likelihoods *B* for the HMM.

Viterbi Algorithm - Pseudocode



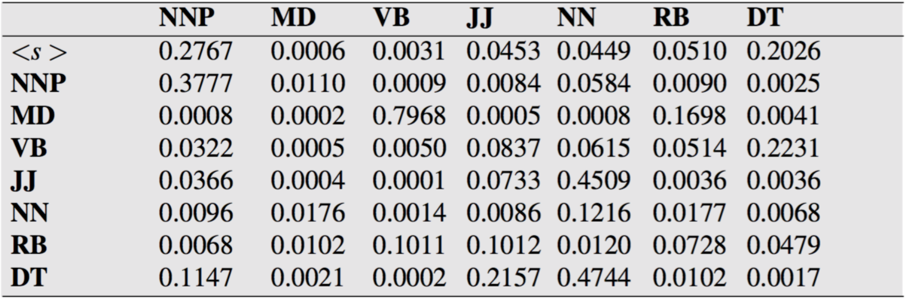
```
function VITERBI(observations of len T,state-graph) returns best-path

  num-states ← NUM-OF-STATES(state-graph)
  Create a path probability matrix viterbi[num-states+2,T+2]
  viterbi[0,0] ← 1.0
  for each time step t from 1 to T do
    for each state s from 1 to num-states do
      viterbi[s,t] ← $\max_{1 \le s' \le num\text{-}states} viterbi[s',t-1] * a_{s',s} * b_s(o_t)$
      backpointer[s,t] ← $\operatorname{argmax}_{1 \le s' \le num\text{-}states} viterbi[s',t-1] * a_{s',s}$
  Backtrace from highest probability state in final column of viterbi[] and return path
```

Figure 6.10 Viterbi algorithm for finding optimal sequence of tags. Given an observation sequence and an HMM $\lambda = (A,B)$, the algorithm returns the state-path through the HMM which assigns maximum likelihood to the observation sequence. Note that states 0 and N+1 are non-emitting *start* and *end* states.

POS Tagging - Example

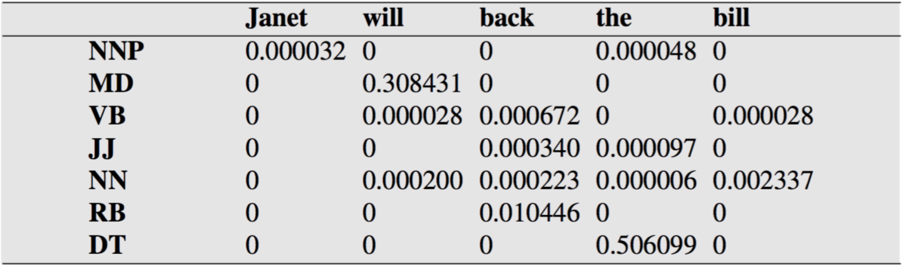
- Janet will back the bill
- Janet/NNP will/MD back/VB the/DT bill/NN



	NNP	MD	VB	JJ	NN	RB	DT
$<s>$	0.2767	0.0006	0.0031	0.0453	0.0449	0.0510	0.2026
NNP	0.3777	0.0110	0.0009	0.0084	0.0584	0.0090	0.0025
MD	0.0008	0.0002	0.7968	0.0005	0.0008	0.1698	0.0041
VB	0.0322	0.0005	0.0050	0.0837	0.0615	0.0514	0.2231
JJ	0.0366	0.0004	0.0001	0.0733	0.4509	0.0036	0.0036
NN	0.0096	0.0176	0.0014	0.0086	0.1216	0.0177	0.0068
RB	0.0068	0.0102	0.1011	0.1012	0.0120	0.0728	0.0479
DT	0.1147	0.0021	0.0002	0.2157	0.4744	0.0102	0.0017

POS Tagging - Example

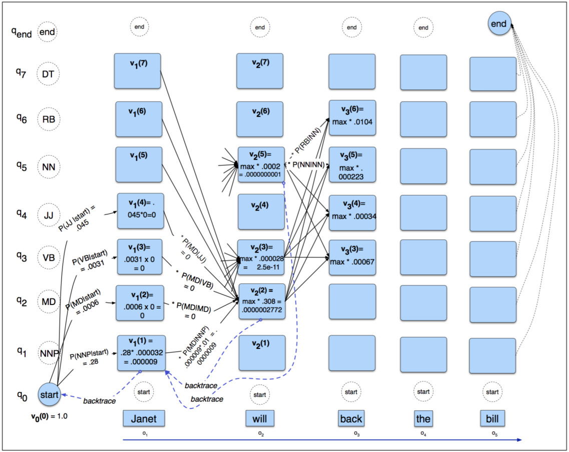
- Janet will back the bill
- Janet/NNP will/MD back/VB the/DT bill/NN



	Janet	**will**	**back**	**the**	**bill**
NNP	0.000032	0	0	0.000048	0
MD	0	0.308431	0	0	0
VB	0	0.000028	0.000672	0	0.000028
JJ	0	0	0.000340	0.000097	0
NN	0	0.000200	0.000223	0.000006	0.002337
RB	0	0	0.010446	0	0
DT	0	0	0	0.506099	0

POS Tagging - Example

- Janet will back the bill
- Janet/NNP will/MD back/VB the/DT bill/NN



Relation Extraction

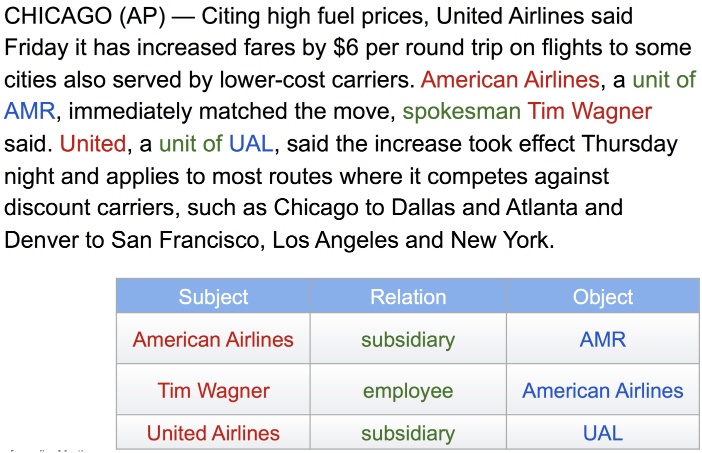
- Relation Extraction
  - Introduction
  - Relation types
- Relation Extraction Methods
  - Hand-built patterns
  - Supervised learning methods
  - Unsupervised or semi-supervised Methods
    - Bootstrapping methods
    - Distant supervision

Relation Extraction - Example

CHICAGO (AP) — Citing high fuel prices, United Airlines said Friday it has increased fares by $6 per round trip on flights to some cities also served by lower-cost carriers. American Airlines, a unit of AMR, immediately matched the move, spokesman Tim Wagner said. United, a unit of UAL, said the increase took effect Thursday night and applies to most routes where it competes against discount carriers, such as Chicago to Dallas and Atlanta and Denver to San Francisco, Los Angeles and New York.

Question: What relations should we extract?

Relation Extraction - Example



CHICAGO (AP) — Citing high fuel prices, United Airlines said Friday it has increased fares by $6 per round trip on flights to some cities also served by lower-cost carriers. American Airlines, a unit of AMR, immediately matched the move, spokesman Tim Wagner said. United, a unit of UAL, said the increase took effect Thursday night and applies to most routes where it competes against discount carriers, such as Chicago to Dallas and Atlanta and Denver to San Francisco, Los Angeles and New York.

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

Relation Types - ACE 2003

ROLE: relates a person to an organization or a geopolitical entity
subtypes: member, owner, affiliate, client, citizen

PART: generalized containment
subtypes: subsidiary, physical part-of, set membership

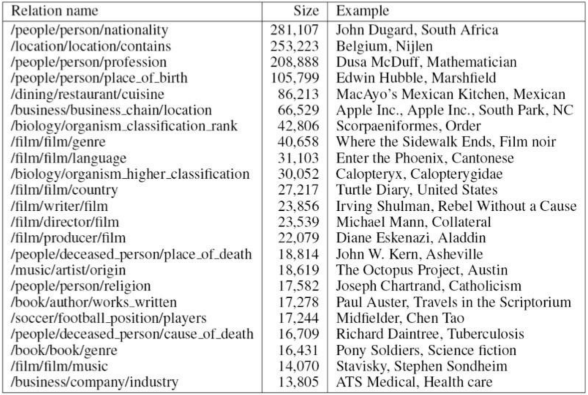
AT: permanent and transient locations
subtypes: located, based-in, residence

SOCIAL: social relations among persons
subtypes: parent, sibling, spouse, grandparent, associate

slide adapted from Doug Appelt

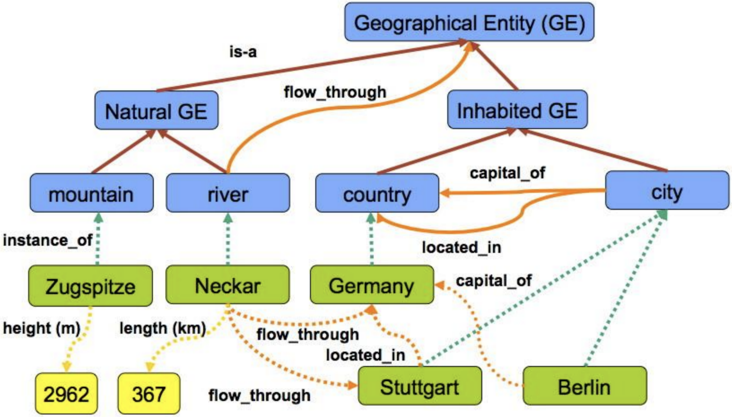
Relation Types - Freebase

23 Million Entities, thousands of relations



Relation name	Size	Example
/people/person/nationality	281,107	John Dugard, South Africa
/location/location/contains	253,223	Belgium, Nijlen
/people/person/profession	208,888	Dusa McDuff, Mathematician
/people/person/place_of_birth	105,799	Edwin Hubble, Marshfield
/dining/restaurant/cuisine	86,213	MacAyo's Mexican Kitchen, Mexican
/business/business_chain/location	66,529	Apple Inc., Apple Inc., South Park, NC
/biology/organism_classification_rank	42,806	Scorpaeniformes, Order
/film/film/genre	40,658	Where the Sidewalk Ends, Film noir
/film/film/language	31,103	Enter the Phoenix, Cantonese
/biology/organism_higher_classification	30,052	Calopteryx, Calopterygidae
/film/film/country	27,217	Turtle Diary, United States
/film/writer/film	23,856	Irving Shulman, Rebel Without a Cause
/film/director/film	23,539	Michael Mann, Collateral
/film/producer/film	22,079	Diane Eskenazi, Aladdin
/people/deceased_person/place_of_death	18,814	John W. Kern, Asheville
/music/artist/origin	18,619	The Octopus Project, Austin
/people/person/religion	17,582	Joseph Chartrand, Catholicism
/book/author/works_written	17,278	Paul Auster, Travels in the Scriptorium
/soccer/football_position/players	17,244	Midfielder, Chen Tao
/people/deceased_person/cause_of_death	16,709	Richard Daintree, Tuberculosis
/book/book/genre	16,431	Pony Soldiers, Science fiction
/film/film/music	14,070	Stavisky, Stephen Sondheim
/business/company/industry	13,805	ATS Medical, Health care

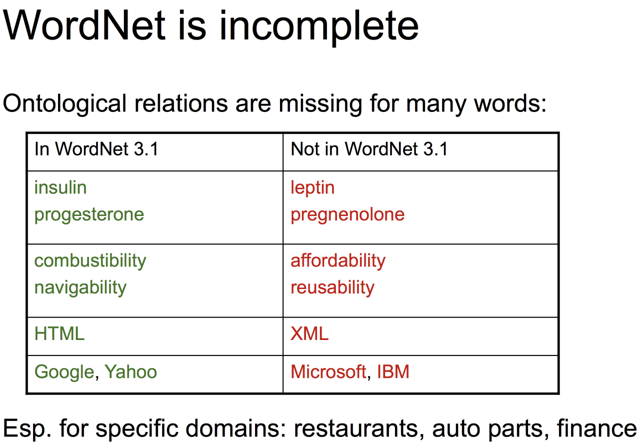
Relation Types - Geospatial



Relation Extraction - Need

- NLP applications need word meaning!
  - Question answering
  - Conversational agents
  - Summarization
- One key meaning component: word relations
  - Hyponymy: San Francisco is an instance of a city
  - Antonymy: acidic is the opposite of basic
  - Meronymy: an alternator is a part of a car

Relation Extraction - Need



WordNet is incomplete

Ontological relations are missing for many words:

In WordNet 3.1	Not in WordNet 3.1
insulin progesterone	leptin pregnenolone
combustibility navigability	affordability reusability
HTML	XML
Google, Yahoo	Microsoft, IBM

Esp. for specific domains: restaurants, auto parts, finance

Hand-built Patterns

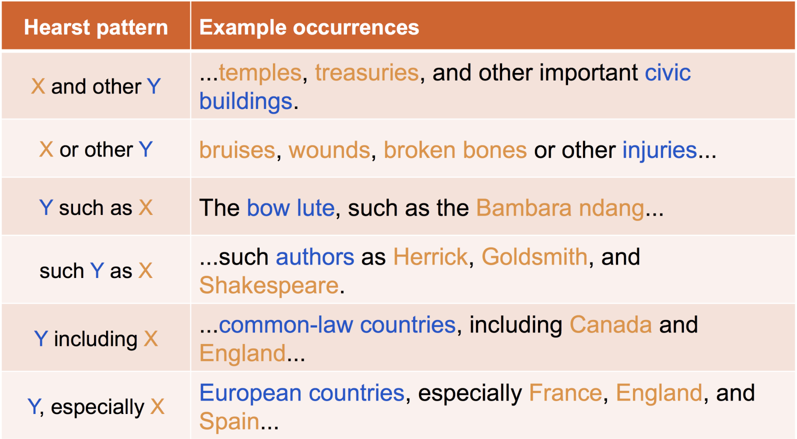
What does *Gelidium* mean?

Hand-built Patterns

What does *Gelidium* mean?

Agar is a substance prepared from a mixture of red algae, such as Gelidium, for laboratory or industrial use

Examples of the Hearst's Patterns



Hearst pattern	Example occurrences
X and other Y	...temples, treasuries, and other important civic buildings.
X or other Y	bruises, wounds, broken bones or other injuries...
Y such as X	The bow lute, such as the Bambara ndang...
such Y as X	...such authors as Herrick, Goldsmith, and Shakespeare.
Y including X	...common-law countries, including Canada and England...
Y, especially X	European countries, especially France, England, and Spain...

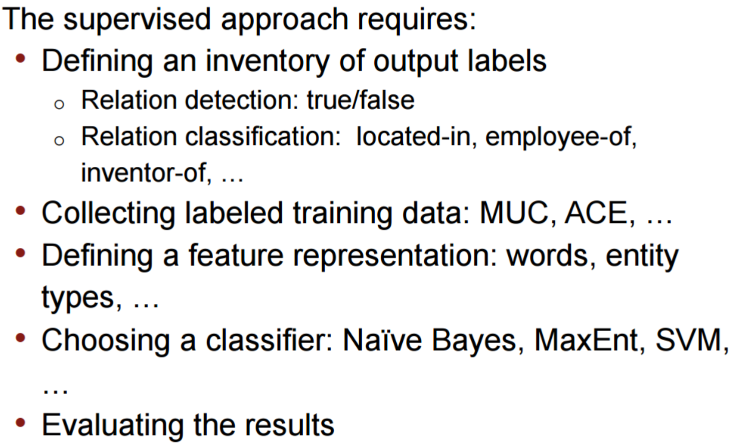
slide adapted from Luke Zettlemoyer

Problems with Hand-built Patterns

- Requires hand-building patterns for each relation!
  - hard to write; hard to maintain
  - there are zillions of them
  - domain-dependent
- Don't want to do this for all possible relations!
- Plus, we'd like better accuracy
  - Hearst: 66% accuracy on hyponym extraction

slide adapted from Luke Zettlemoyer

Supervised Relation Extraction - Approach

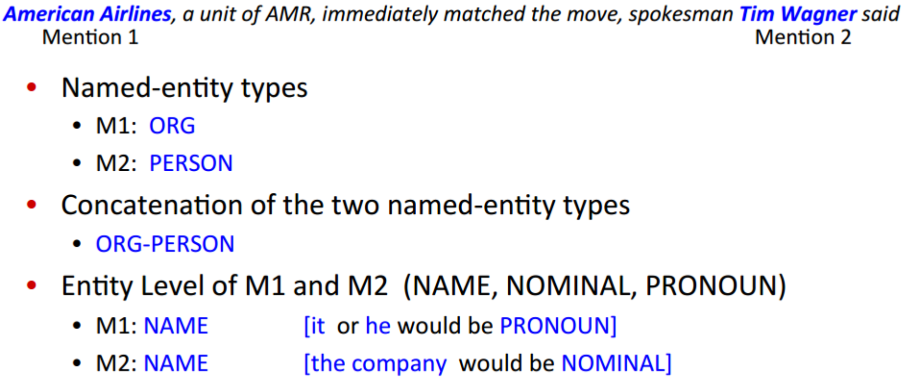


The supervised approach requires:

- Defining an inventory of output labels
  - Relation detection: true/false
  - Relation classification: located-in, employee-of, inventor-of, …
- Collecting labeled training data: MUC, ACE, …
- Defining a feature representation: words, entity types, …
- Choosing a classifier: Naïve Bayes, MaxEnt, SVM, …
- Evaluating the results

slide adapted from Luke Zettlemoyer

Word Features for Relation Extraction



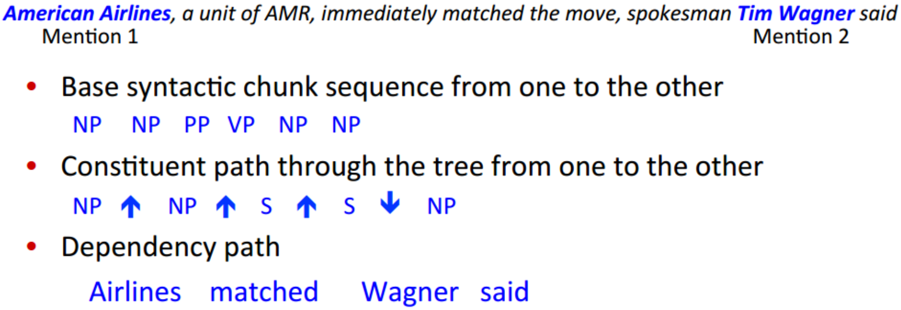
***American Airlines**, a unit of AMR, immediately matched the move, spokesman **Tim Wagner** said*

Mention 1 Mention 2

- Named-entity types
  - M1: ORG
  - M2: PERSON
- Concatenation of the two named-entity types
  - ORG-PERSON
- Entity Level of M1 and M2 (NAME, NOMINAL, PRONOUN)
  - M1: NAME [it or he would be PRONOUN]
  - M2: NAME [the company would be NOMINAL]

slide adapted from Dan Jurafsky

Parse Features for Relation Extraction



American Airlines*, a unit of AMR, immediately matched the move, spokesman* ***Tim Wagner*** *said*

Mention 1 Mention 2

- Base syntactic chunk sequence from one to the other

  NP NP PP VP NP NP

- Constituent path through the tree from one to the other

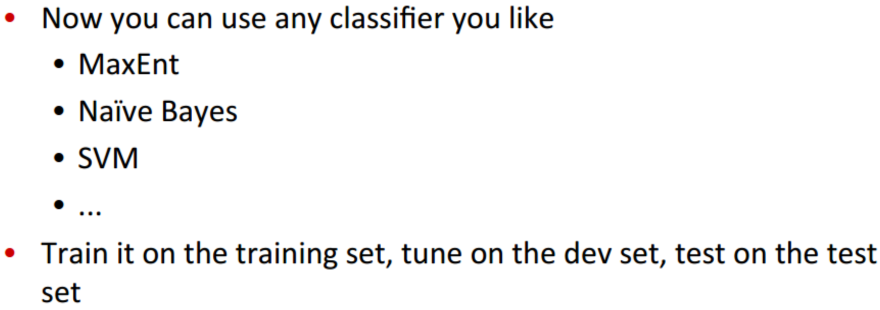
  NP ↑ NP ↑ S ↑ S ↓ NP

- Dependency path

  Airlines matched Wagner said

slide adapted from Dan Jurafsky

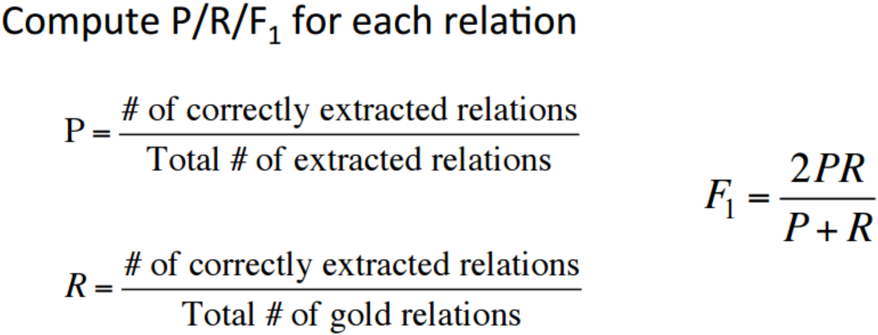
Classifiers for Supervised Learning Methods



- Now you can use any classifier you like
  - MaxEnt
  - Naïve Bayes
  - SVM
  - ...
- Train it on the training set, tune on the dev set, test on the test set

slide adapted from Dan Jurafsky

Supervised Learning Methods - Evaluation



Compute P/R/F$_1$ for each relation

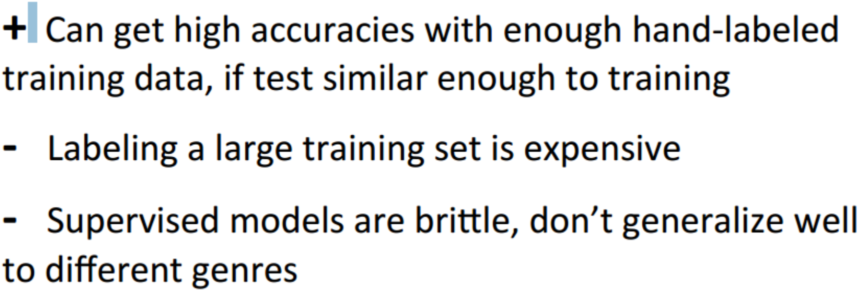
$$P = \frac{\text{\# of correctly extracted relations}}{\text{Total \# of extracted relations}}$$

$$R = \frac{\text{\# of correctly extracted relations}}{\text{Total \# of gold relations}}$$

$$F_1 = \frac{2PR}{P+R}$$

slide adapted from Dan Jurafsky

Supervised Learning Methods - Summary



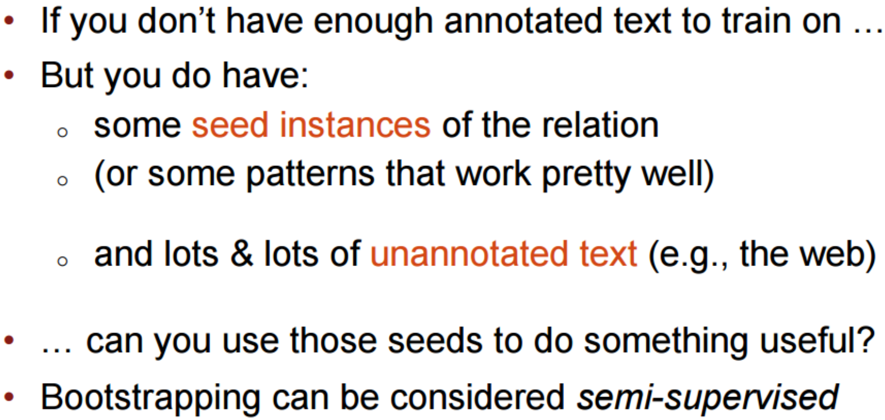
+ Can get high accuracies with enough hand-labeled training data, if test similar enough to training

- Labeling a large training set is expensive

- Supervised models are brittle, don’t generalize well to different genres

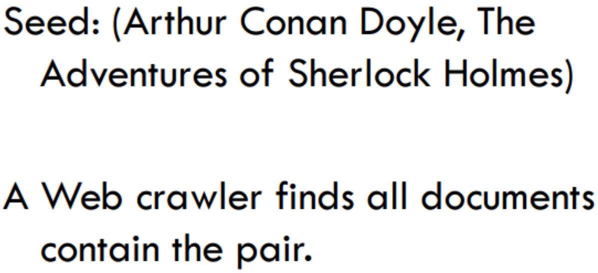
slide adapted from Dan Jurafsky

Bootstrapping Approach



- If you don't have enough annotated text to train on …
- But you do have:
  - some seed instances of the relation
  - (or some patterns that work pretty well)
  - and lots & lots of unannotated text (e.g., the web)
- … can you use those seeds to do something useful?
- Bootstrapping can be considered *semi-supervised*

Bootstrapping Example

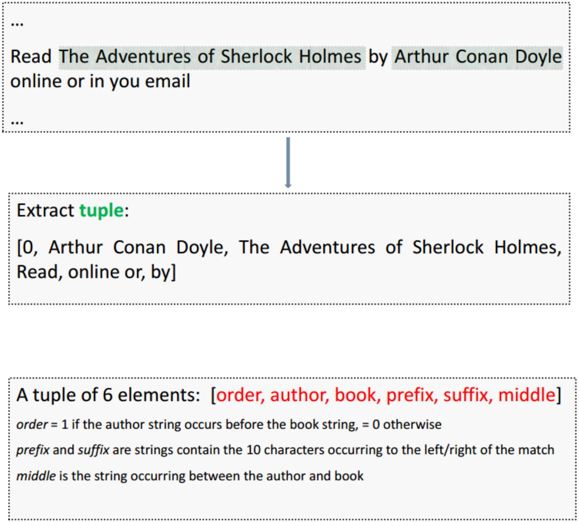


Seed: (Arthur Conan Doyle, The Adventures of Sherlock Holmes)

A Web crawler finds all documents contain the pair.

slide adapted from Nguyen Bach and Sameer Badaskar

Bootstrapping - Matched Document 1



...

Read The Adventures of Sherlock Holmes by Arthur Conan Doyle online or in you email

...

Extract **tuple**:

[0, Arthur Conan Doyle, The Adventures of Sherlock Holmes, Read, online or, by]

A tuple of 6 elements: [order, author, book, prefix, suffix, middle]

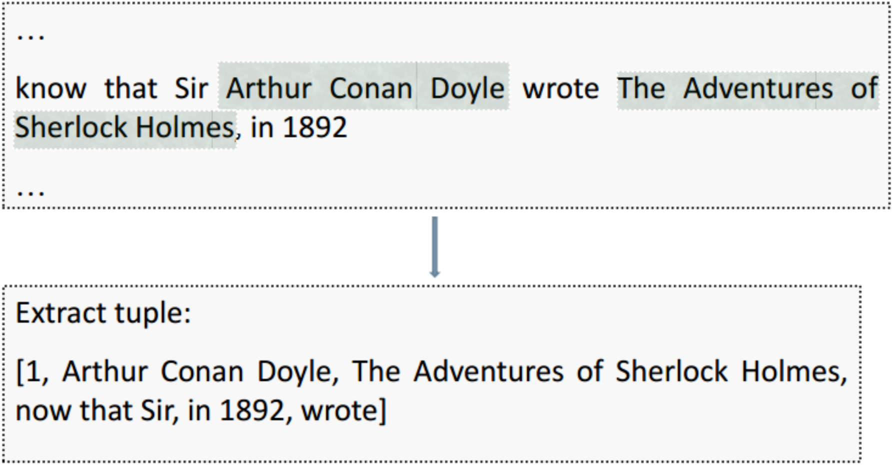
order = 1 if the author string occurs before the book string, = 0 otherwise

prefix and *suffix* are strings contain the 10 characters occurring to the left/right of the match

middle is the string occurring between the author and book

slide adapted from Nguyen Bach and Sameer Badaskar

Bootstrapping - Matched Document 2



...

know that Sir Arthur Conan Doyle wrote The Adventures of Sherlock Holmes, in 1892

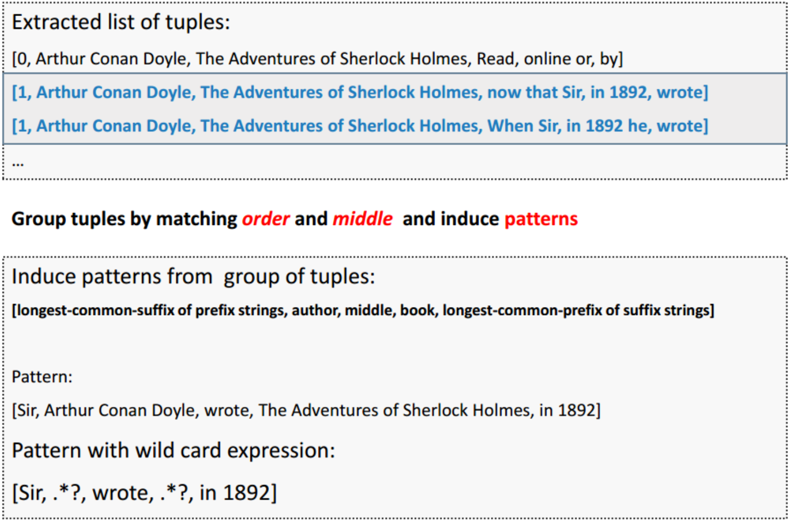
...

Extract tuple:

[1, Arthur Conan Doyle, The Adventures of Sherlock Holmes, now that Sir, in 1892, wrote]

slide adapted from Nguyen Bach and Sameer Badaskar

Bootstrapping - Developing Patterns



Extracted list of tuples:

[0, Arthur Conan Doyle, The Adventures of Sherlock Holmes, Read, online or, by]

[1, Arthur Conan Doyle, The Adventures of Sherlock Holmes, now that Sir, in 1892, wrote]

[1, Arthur Conan Doyle, The Adventures of Sherlock Holmes, When Sir, in 1892 he, wrote]

…

Group tuples by matching *order* and *middle* and induce patterns

Induce patterns from group of tuples:

[longest-common-suffix of prefix strings, author, middle, book, longest-common-prefix of suffix strings]

Pattern:

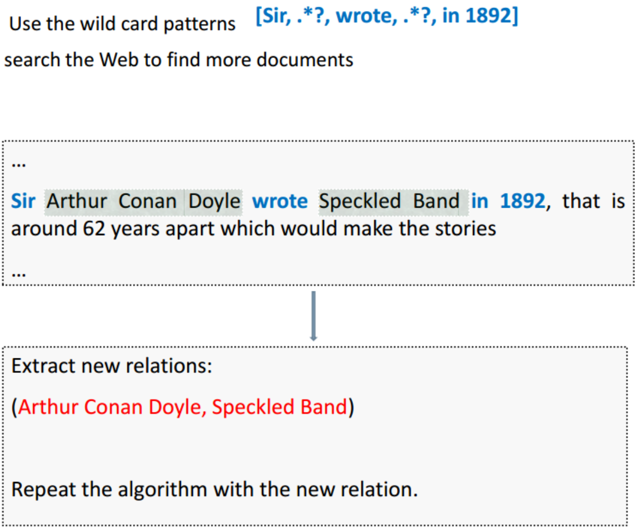
[Sir, Arthur Conan Doyle, wrote, The Adventures of Sherlock Holmes, in 1892]

Pattern with wild card expression:

[Sir, .*?, wrote, .*?, in 1892]

slide adapted from Nguyen Bach and Sameer Badaskar

Bootstrapping - Search for more patterns and new tuples



Use the wild card patterns [Sir, .*?, wrote, .*?, in 1892]

search the Web to find more documents

...

Sir Arthur Conan Doyle wrote Speckled Band in 1892, that is around 62 years apart which would make the stories

...

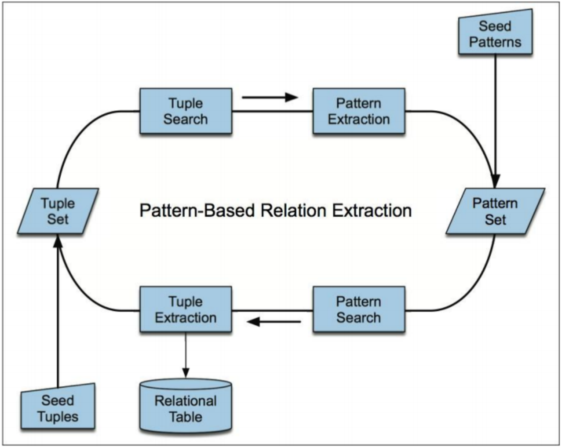
Extract new relations:

(Arthur Conan Doyle, Speckled Band)

Repeat the algorithm with the new relation.

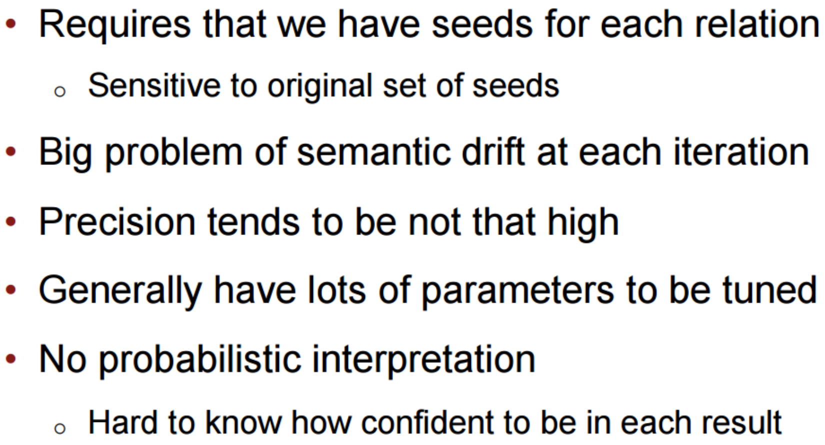
slide adapted from Nguyen Bach and Sameer Badaskar

Bootstrapping System



slide adapted from Jim Martin

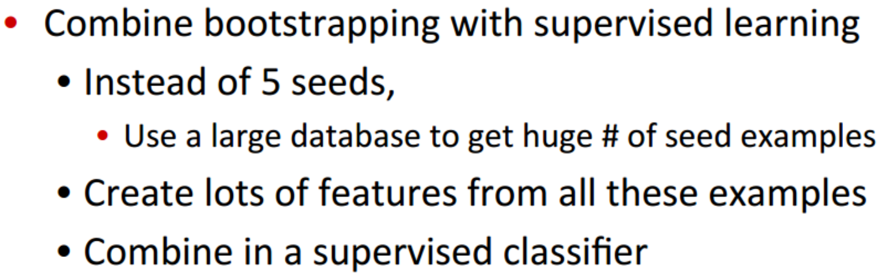
Bootstrapping - Problems



- Requires that we have seeds for each relation
  - Sensitive to original set of seeds
- Big problem of semantic drift at each iteration
- Precision tends to be not that high
- Generally have lots of parameters to be tuned
- No probabilistic interpretation
  - Hard to know how confident to be in each result

slide adapted from Luke Zettlemoyer

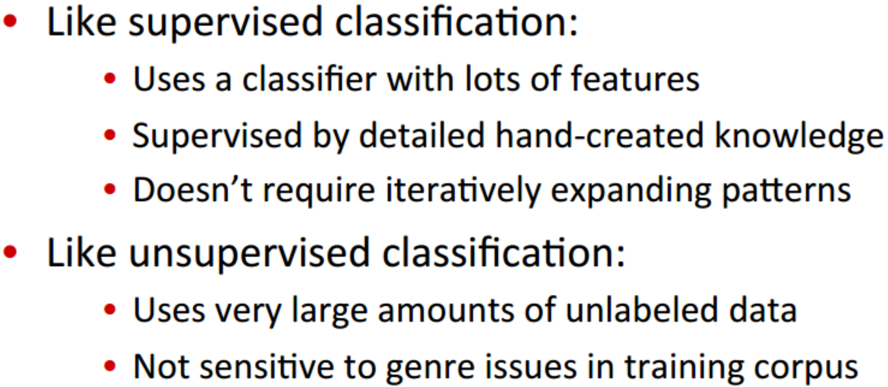
Distant supervision method



- Combine bootstrapping with supervised learning
  - Instead of 5 seeds,
    - Use a large database to get huge # of seed examples
  - Create lots of features from all these examples
  - Combine in a supervised classifier

slide adapted from Dan Jurafsky

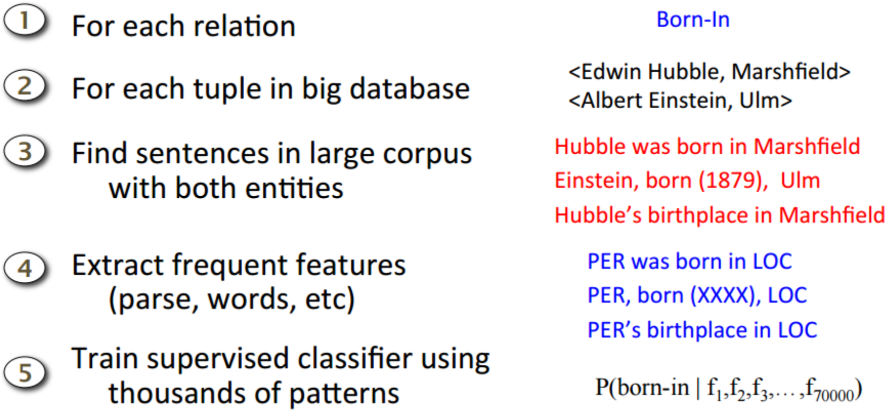
Distant supervision paragidm



- Like supervised classification:
  - Uses a classifier with lots of features
  - Supervised by detailed hand-created knowledge
  - Doesn’t require iteratively expanding patterns
- Like unsupervised classification:
  - Uses very large amounts of unlabeled data
  - Not sensitive to genre issues in training corpus

slide adapted from Dan Jurafsky

Distant supervision - approach



1. For each relation
2. For each tuple in big database
3. Find sentences in large corpus with both entities
4. Extract frequent features (parse, words, etc)
5. Train supervised classifier using thousands of patterns

Born-In

<Edwin Hubble, Marshfield>
<Albert Einstein, Ulm>

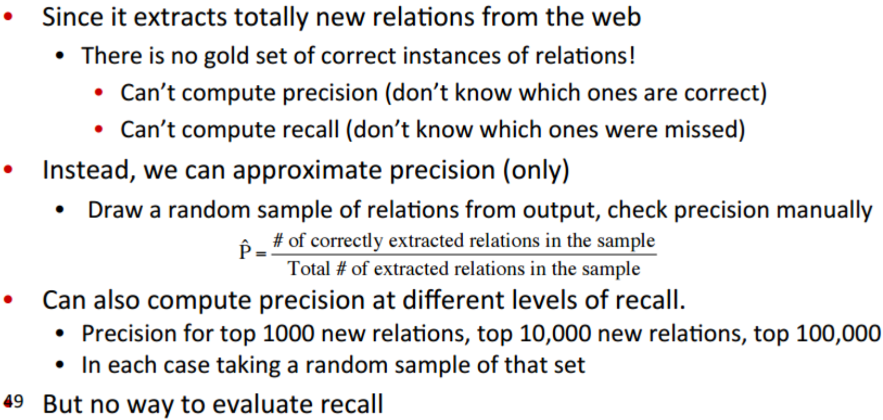
Hubble was born in Marshfield
Einstein, born (1879), Ulm
Hubble's birthplace in Marshfield

PER was born in LOC
PER, born (XXXX), LOC
PER's birthplace in LOC

$P(\text{born-in} \mid f_1, f_2, f_3, \ldots, f_{70000})$

slide adapted from Dan Jurafsky

Distant supervision - approach



- Since it extracts totally new relations from the web
  - There is no gold set of correct instances of relations!
    - Can't compute precision (don't know which ones are correct)
    - Can't compute recall (don't know which ones were missed)
- Instead, we can approximate precision (only)
  - Draw a random sample of relations from output, check precision manually

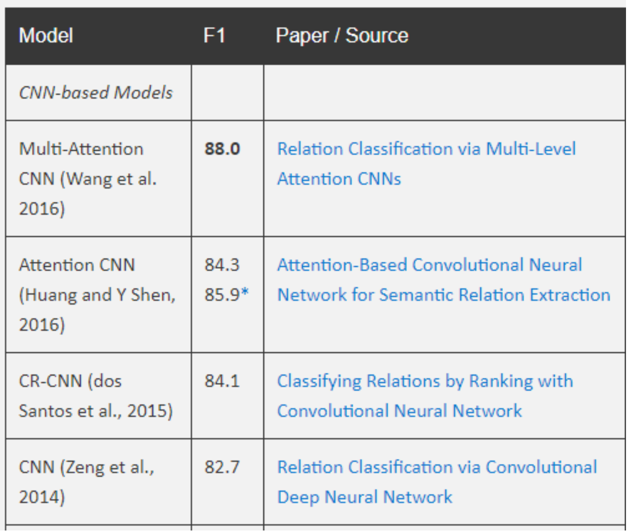
$$\hat{P} = \frac{\text{\# of correctly extracted relations in the sample}}{\text{Total \# of extracted relations in the sample}}$$

- Can also compute precision at different levels of recall.
  - Precision for top 1000 new relations, top 10,000 new relations, top 100,000
  - In each case taking a random sample of that set

49 But no way to evaluate recall

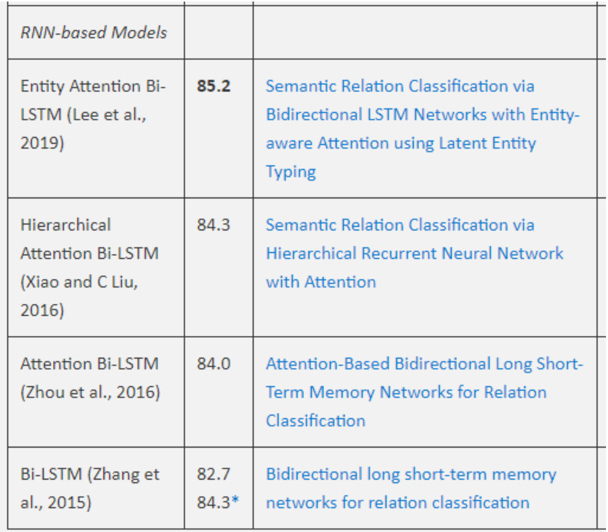
slide adapted from Dan Jurafsky

CNN-based Models



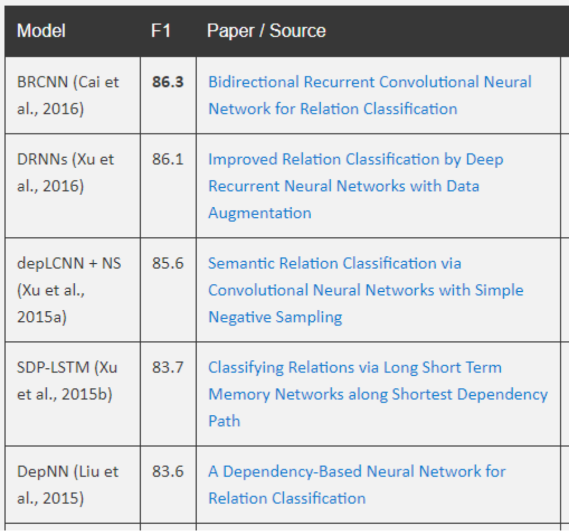
Model	F1	Paper / Source
CNN-based Models		
Multi-Attention CNN (Wang et al. 2016)	**88.0**	Relation Classification via Multi-Level Attention CNNs
Attention CNN (Huang and Y Shen, 2016)	84.3 85.9*	Attention-Based Convolutional Neural Network for Semantic Relation Extraction
CR-CNN (dos Santos et al., 2015)	84.1	Classifying Relations by Ranking with Convolutional Neural Network
CNN (Zeng et al., 2014)	82.7	Relation Classification via Convolutional Deep Neural Network

RNN-based Models



RNN-based Models		
Entity Attention Bi-LSTM (Lee et al., 2019)	**85.2**	Semantic Relation Classification via Bidirectional LSTM Networks with Entity-aware Attention using Latent Entity Typing
Hierarchical Attention Bi-LSTM (Xiao and C Liu, 2016)	84.3	Semantic Relation Classification via Hierarchical Recurrent Neural Network with Attention
Attention Bi-LSTM (Zhou et al., 2016)	84.0	Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification
Bi-LSTM (Zhang et al., 2015)	82.7 84.3*	Bidirectional long short-term memory networks for relation classification

Dependency-based NN Models



Model	F1	Paper / Source
BRCNN (Cai et al., 2016)	**86.3**	Bidirectional Recurrent Convolutional Neural Network for Relation Classification
DRNNs (Xu et al., 2016)	86.1	Improved Relation Classification by Deep Recurrent Neural Networks with Data Augmentation
depLCNN + NS (Xu et al., 2015a)	85.6	Semantic Relation Classification via Convolutional Neural Networks with Simple Negative Sampling
SDP-LSTM (Xu et al., 2015b)	83.7	Classifying Relations via Long Short Term Memory Networks along Shortest Dependency Path
DepNN (Liu et al., 2015)	83.6	A Dependency-Based Neural Network for Relation Classification