Information Extraction

Instructor: Thien Huu Nguyen

Based on slides from: Ralph Grishman

Information Extraction (IE)

Giuliani, 58, proposed to Nathan, a former nurse, during a business trip to Paris _ five months after he finalized his divorce from Donna Hanover in July after 20 years of marriage.

In interviews last year, Giuliani said Nathan gave him ``tremendous emotional support'' through his treatment for prostate cancer and as he led New York City during the Sept. 11, 2001, terror attacks.

Relation Knowledge Base

Name	leaderOf	....
Giuliani	New York City	
.....		

Data Mining

Reasoning

Monitoring

...

IE = automatically extracting structured information from unstructured and/or semi-structured machine-readable documents

Event Knowledge Base

Trigger	Type	Person1	Person2	Time
divorce	Divorce	Giuliani	Donna Hanover	July
.....				

Information Extraction Pipeline

Giuliani, 58, proposed to Nathan, a former nurse, during a business trip to Paris _ five months after he finalized his divorce from Donna Hanover in July after 20 years of marriage.

In interviews last year, Giuliani said Nathan gave him ``tremendous emotional support'' through his treatment for prostate cancer and as he led New York City during the Sept. 11, 2001, terror attacks.

Relation Knowledge Base

Corpora

Event Knowledge Base

Information Extraction Pipeline

Person

Location

Time

Giuliani, 58, proposed to Nathan, a former nurse, during a business trip to Paris _ five months after he finalized his divorce from Donna Hanover in July after 20 years of marriage.

In interviews last year, Giuliani said Nathan gave him ``tremendous emotional support'' through his treatment for prostate cancer and as he led New York City during the Sept. 11, 2001, terror attacks.

Relation Knowledge Base

Name ....

Giuliani

.....

Corpora

Entity Recognition

Event Knowledge Base

Information Extraction Pipeline

Person

Location

Time

Giuliani, 58, proposed to Nathan, a former nurse, during a business trip to Paris _ five months after he finalized his divorce from Donna Hanover in July after 20 years of marriage.

In interviews last year, Giuliani said Nathan gave him ``tremendous emotional support'' through his treatment for prostate cancer and as he led New York City during the Sept. 11, 2001, terror attacks.

leaderOf

Relation Knowledge Base

Name	....
Giuliani	
.....	

Corpora

Entity Recognition

Relation Extraction

Event Knowledge Base

Information Extraction Pipeline

Person

Location

Time

Giuliani, 58, proposed to Nathan, a former nurse, during a business trip to Paris _ five months after he finalized his divorce from Donna Hanover in July after 20 years of marriage.

In interviews last year, Giuliani said Nathan gave him ``tremendous emotional support'' through his treatment for prostate cancer and as he led New York City during the Sept. 11, 2001, terror attacks.

Corefered

Relation Knowledge Base

Name ....

Giuliani

.....

leaderOf

Corpora

Entity Recognition

Relation Extraction

Coreference Resolution

Event Knowledge Base

Information Extraction Pipeline

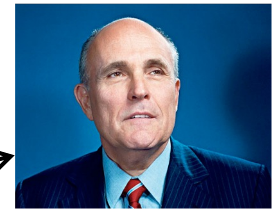
Person

Location

Time

Giuliani, 58, proposed to Nathan, a former nurse, during a business trip to Paris _ five months after he finalized his divorce from Donna Hanover in July after 20 years of marriage.

In interviews last year, Giuliani said Nathan gave him ``tremendous emotional support'' through his treatment for prostate cancer and as he led New York City during the Sept. 11, 2001, terror attacks.



Relation Knowledge Base

Corefered

leaderOf

Name	....
Giuliani	
.....	

Corpora

Entity Recognition

Relation Extraction

Coreference Resolution

Entity Linking

Event Knowledge Base

Information Extraction Pipeline

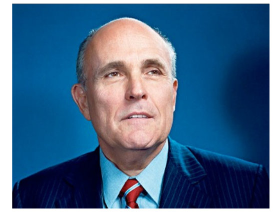
Person

Location

Time

Giuliani, 58, proposed to Nathan, a former nurse, during a business trip to Paris _ five months after he finalized his divorce from Donna Hanover in July after 20 years of marriage.

In interviews last year, Giuliani said Nathan gave him ``tremendous emotional support'' through his treatment for prostate cancer and as he led New York City during the Sept. 11, 2001, terror attacks.



Relation Knowledge Base

Name	leaderOf	....
Giuliani	New York City	
.....		

Corpora

Entity Recognition

Relation Extraction

Coreference Resolution

Entity Linking

Event Knowledge Base

Information Extraction Pipeline

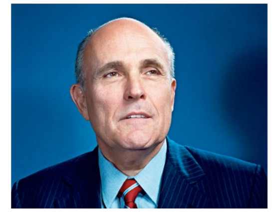
Person

Location

Time

Giuliani, 58, proposed to Nathan, a former nurse, during a business trip to Paris _ five months after he finalized his divorce from Donna Hanover in July after 20 years of marriage.

In interviews last year, Giuliani said Nathan gave him ``tremendous emotional support'' through his treatment for prostate cancer and as he led New York City during the Sept. 11, 2001, terror attacks.



Relation Knowledge Base

Name	leaderOf	....
Giuliani	New York City	
.....		

Corpora

Entity Recognition

Relation Extraction

Coreference Resolution

Entity Linking

Event Knowledge Base

Trigger Prediction

Trigger	Type	Person1	Person2	Time
divorce	Divorce			
.....				

Information Extraction Pipeline

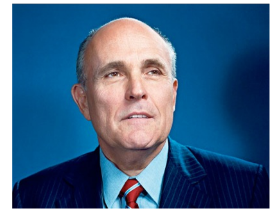
Person

Location

Time

Giuliani, 58, proposed to Nathan, a former nurse, during a business trip to Paris _ five months after he finalized his divorce from Donna Hanover in July after 20 years of marriage.

In interviews last year, Giuliani said Nathan gave him ``tremendous emotional support'' through his treatment for prostate cancer and as he led New York City during the Sept. 11, 2001, terror attacks.



Relation Knowledge Base

Name	leaderOf	....
Giuliani	New York City	
.....		

Corpora

Entity Recognition

Relation Extraction

Coreference Resolution

Entity Linking

Event Knowledge Base

Trigger Prediction

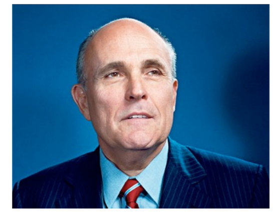
Argument Prediction

Trigger	Type	Person1	Person2	Time
divorce	Divorce			
.....				

Information Extraction Pipeline

Giuliani, 58, proposed to Nathan, a former nurse, during a business trip to Paris _ five months after he finalized his divorce from Donna Hanover in July after 20 years of marriage.

In interviews last year, Giuliani said Nathan gave him ``tremendous emotional support'' through his treatment for prostate cancer and as he led New York City during the Sept. 11, 2001, terror attacks.



Relation Knowledge Base

Name	leaderOf	....
Giuliani	New York City	
.....		

Corpora

Entity Recognition

Relation Extraction

Coreference Resolution

Entity Linking

Event Knowledge Base

Trigger Prediction

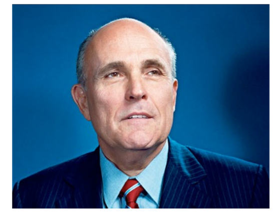
Argument Prediction

Trigger	Type	Person1	Person2	Time
divorce	Divorce			
.....				

Information Extraction Pipeline

Giuliani, 58, proposed to Nathan, a former nurse, during a business trip to Paris _ five months after he finalized his divorce from Donna Hanover in July after 20 years of marriage.

In interviews last year, Giuliani said Nathan gave him ``tremendous emotional support'' through his treatment for prostate cancer and as he led New York City during the Sept. 11, 2001, terror attacks.



Relation Knowledge Base

Name	leaderOf	....
Giuliani	New York City	
.....		

Corpora

Entity Recognition

Relation Extraction

Coreference Resolution

Entity Linking

Event Knowledge Base

Trigger Prediction

Argument Prediction

Trigger	Type	Person1	Person2	Time
divorce	Divorce	Giuliani	Donna Hanover	July
.....				

Information Extraction vs. Information Retrieval

- Information Retrieval returns a set of documents given a query.
- Information Extraction returns facts from documents
- E.g., What you search for in real estate advertisements:
  - Town/suburd. You might think easy, but:
    - Real estate agents: Coldwell Banker, Mosman
    - Phrases: Only 45 minutes from Parramatta
    - Multiple property ads have different suburbs in one ad
  - Money: want a range not a textual match
    - Multiple amounts: was $155K, now $145K
  - Bedrooms
    - Variations: br, bdr, beds, B/R

Information Extraction Evaluations

- CoNLL has sponsored annual evaluations of NLP components for about 15 years

- NIST has organized (annual) US Government evaluations of information extraction for about 25 years
  - covering both components and integrated systems
  - MUC [Message Understanding Conferences] in the 1990's
  - ACE [Automatic Content Extraction] 2000-2008
  - KBP [Knowledge Base Population] since 2009

Supervised learning for NER

- Named entities are crucial to different IE and QA tasks
- For Named Entity Recognition (NER) (find and classify names in text) , we can use the sequence labeling methods discussed previously (i.e., MEMM, CRF, RNN).

Person — Organization

Fred	Smith	works	for	Time	inc.
B_PER	I_PER	O	O	B_ORG	B_ORG

- Feature-based models: the key is to design good feature sets to feed into the sequence labeling models (i.e., feature engineering with MEMM or CRF)

Features for NER

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

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

prefix(w_i) = `L`	suffix(w_i) = `tane`
prefix(w_i) = `L'`	suffix(w_i) = `ane`
prefix(w_i) = `L'O`	suffix(w_i) = `ne`
prefix(w_i) = `L'Oc`	suffix(w_i) = `e`
word-shape(w_i) = `X'Xxxxxxxx`	short-word-shape(w_i) = `X'Xx`

Features for NER

- Word shape features: Map words to simplified representation that encodes attributes such as length, capitalization, numerals, Greek letters, internal punctuation, etc.

Varicella-zoster	Xx-xxx
mRNA	xXXX
CPA1	XXXd

- Shorter word shape features: consecutive character types are removed (i.e., DC10-30 -> Xd-d, I.M.F -> X.X.X)
- Gazetteers: Lists of common names for different types
  - Millions of entries for locations with detailed geographical and political information (www.geonames.org)
  - Lists of first names and surnames derived from its decadal census in the U.S (www.census.gov)
  - Typically implemented as a binary feature for each name list
  - Unfortunately, such lists can be difficult to create and maintain, and their usefulness varies considerably.

Deep learning for NER

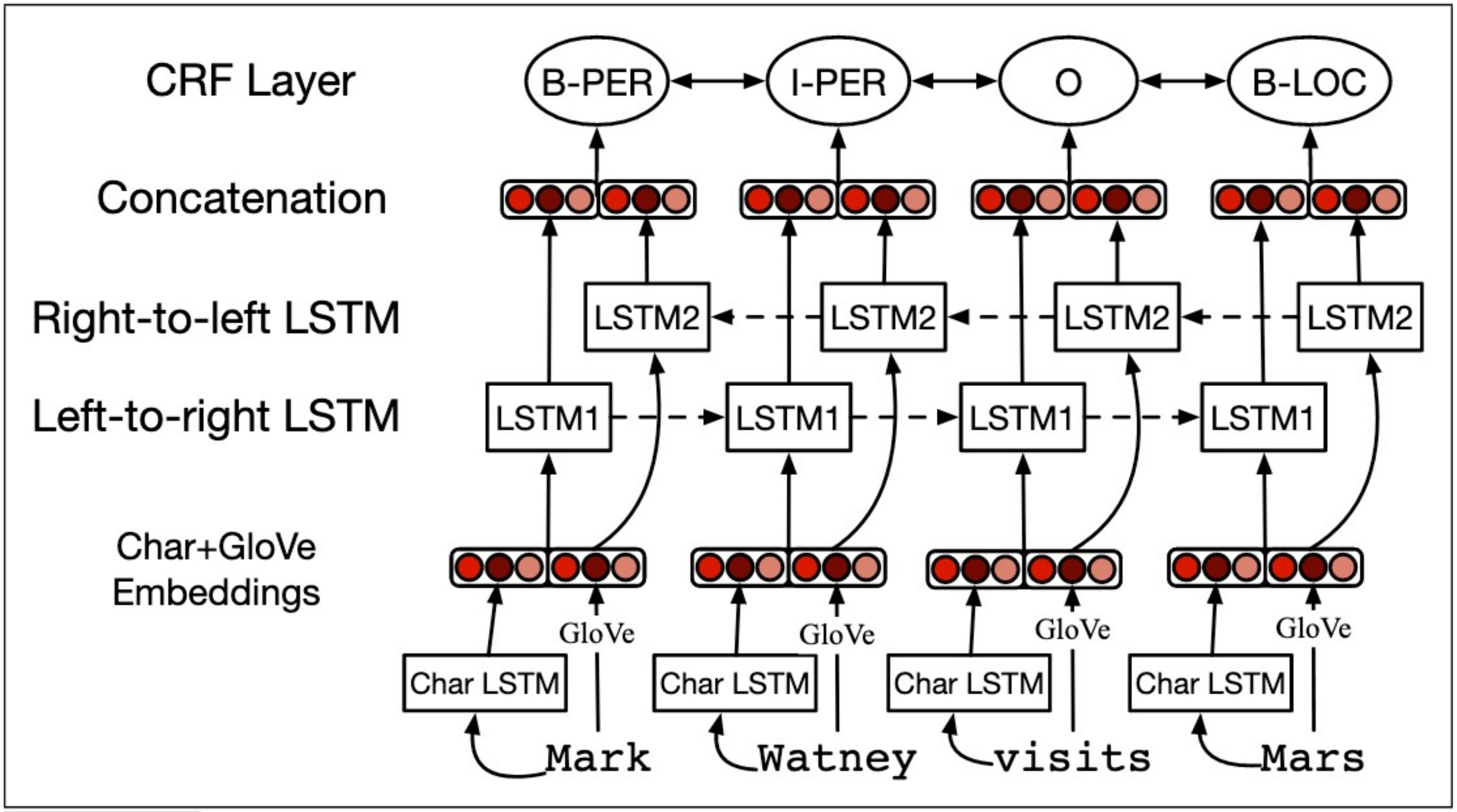


Figure 17.8 Putting it all together: character embeddings and words together a bi-LSTM sequence model. After Lample et al. (2016).

Evaluation for NER systems

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

<start, end, type>

Precision	1/3
Recall	1/2

gold

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

system

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

Supervised learning for Relation Extraction

- A *relation* is a predication about a pair of entities:
  - Rodrigo works for UNED.
  - Alfonso lives in Tarragona.
  - Otto’s father is Ferdinand.
- Typically they represent information which is permanent or of extended duration.

History of relations

- Relations were introduced in MUC-7 (1997)
  - 3 relations
- Extensively studied in ACE (2000 – 2007)
  - lots of training data
- Effectively included in KBP
  - Wikipedia infobox model

ACE Relations

- Several revisions of relation definitions
    - With goal of having a set of relations which can be consistently annotated
- 5-7 major types, 19-24 subtypes
- Both entities must be mentioned in the same sentence
    - Do not get a parent-child relation from
        - Ferdinand and Isabella were married in 1481. A son was born in 1485.
    - Or an employee relation for
        - Bank Santander replaced several executives. Alfonso was named an executive vice president.
- Base for extensive research
    - On supervised and semi-supervised methods

2004 Ace Relation Types

Relation type	Subtypes
Physical	Located, Near, Part-whole
Personal-social	Business, Family, Other
Employment / Membership / Subsidiary	Employ-executive, Employ-staff, Employ-undetermined, Member-of-group, Partner, Subsidiary, Other
Agent-artifact	User-or-owner, Inventor-or-manufacturer, Other
Person-org affiliation	Ethnic, Ideology, Other
GPE affiliation	Citizen-or-resident, Based-in, Other
Discourse	-

KBP Slots

- Many KBP slots represent relations between entities:
  - Member_of
  - Employee_of
  - Country_of_birth
  - Countries_of_residence
  - Schools_attended
  - Spouse
  - Parents
  - Children …
- Entities do not need to appear in the same sentence
- More limited training data
  - Encouraged semi-supervised methods

Characteristics of Relations

- Relations appear in a wide range of forms:
  - Embedded constructs (one argument contains the other)
    - within a single noun group
      - John's wife
    - linked by a preposition
      - the president of Apple
  - Formulaic constructs
      - Tarragona, Spain
      - Walter Cronkite, CBS News, New York
  - Longer-range ('predicate-linked') constructs
    - With a predicate disjoint from the arguments
      - Fred lived in New York
      - Fred and Mary got married

Methods for Relation Extraction (RE)

- Rule-based methods
  - Write rules to capture different types of relations
- Feature-based methods
  - Design feature sets for RE and send them to some statistical classifiers (i.e., MaxEnt, SVM)
- Kernel-based methods
  - Design kernels to compute similarities between pairs of entities and use them in kernel-based SVM
- Deep learning methods
  - Let deep learning learn the features for RE from data

Rule-based methods for RE: Hand-crafted patterns

- Most instances of relations can be identified by the types of the entities and the words between the entities
  - But not all: Fred and Mary got married.
- Word sequence patterns work well enough for short-range relations
  - But problems arise for longer-range patterns … greater variety, intervening modifiers

Parsing

- progress through corpus-trained parsers
  - probabilistic context-free parsers
  - corpus-trained shift-reduce parsers
  - more accurate, much faster
- how do we take advantage of parsing?
  - arguments of semantic relation generally connected by a limited set of syntactic structures and lexical items
  - need not take into account the wide range of intervening words

Parsing

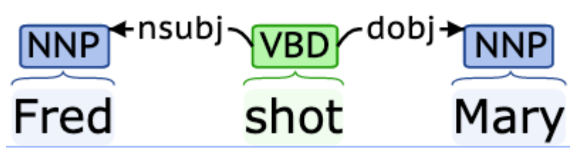
- "Fred shot Mary."
- "Fred, 61, shot Mary."
- "Fred, tired of her endless lectures on parsing, shot Mary."

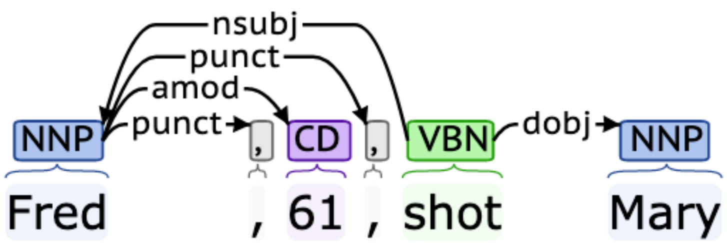
- all have the same dependency relations:
  - verb "shot"
  - subject of shot = "Fred"
  - object of shot = "Mary"

Lexicalized Dependency Paths

- using path in dependency tree between two entity mentions as the patterns for RE
- combines dependency types and lexical items
  - type = edge from governor to dependent
  - type-1 = edge from dependent to governor

PERSON – nsubj-1:shot:dobj -- PERSON





Supervised learning for RE

- Collect training data
  - Annotate corpus with entities and relations
  - For every pair of entities in a sentence
    - If linked by a relation, treat as positive training instance
    - If not linked, treat as a negative training instance
- Train model
  - For *n* relation types, either
    - Binary (identification) model + *n*-way classifier model or
    - Unified *n+1*-way classifier
    - Either way, the dataset is very imbalanced toward the negative instances (“Other”)
- On test data
  - Apply entity classifier
  - Apply relation classifier to every pair of entities in same sentence
- Evaluate using Precision, Recall and F1

Supervised learning for RE

- The spokesman, reporting on the meeting, said IBM hired Fred Smith as the president.

Relation instances

- The spokesman, reporting on the meeting, said IBM hired Fred Smith as the president. -> Other
- The spokesman, reporting on the meeting, said IBM hired Fred Smith as the president. -> Other
- The spokesman, reporting on the meeting, said IBM hired Fred Smith as the president. -> Other
- The spokesman, reporting on the meeting, said IBM hired Fred Smith as the president. -> Employment
- The spokesman, reporting on the meeting, said IBM hired Fred Smith as the president. -> Employment
- The spokesman, reporting on the meeting, said IBM hired Fred Smith as the president. -> Other

Feature-based methods for RE

- Design a set of features, compute the values of such features for each instance, and send them statistical classifiers for classification

- Typical features:
  - Heads of entities
  - Types of entities
  - Distance between entities
  - Containment relations
  - Word sequence between entities
  - Individual words between entities
  - Dependency path
  - Individual words on dependency path

Zhou et al., 2005: Exploring Various Knowledge in Relation Extraction (ACL)

Features for RE

Ray Young, the chief financial officer of General Motors, said GM could not bail out Delphi

Designed Features	Values	Designed Features	Values
head word of M1	Ray_Young	last word in between	of
head word of M2	General_ Motors	middle token sequence	, the chief financial officer of
first word before M1	nil	Shortest path connecting M1 and M2 in the dependency parsing tree	PERSON_appos_officer prep_of_ORGANIZATION
second word before M1	nil		
first word after M2	,	entity type of M1	PERSON
second word after M2	said	enity type of M2	ORGANIZATION
first word in between	,	...	….


```
BagWM1_mark=TRUE BagWM1_webster=TRUE BagWM2_itn=TRUE WBFL=of BM1F=first. BM1L=at AM2F=has AM2L=an NUMWB=1 TPatternET=PERSON_of_ORGANIZATION PTP=NP----PP PTPH=NP----NP--webster----PP CPHBFL=of CPP=NP----PP----NP CPPH=NP--webster-
---PP--of----NP--itn CPatternET=PERSON_of_ORGANIZATION CPHAM2F=has CPHAM2L=update DPathET=PERSON_prep_of_ORGANIZATION ET1DW1=PERSON--has ET2DW2=ORGANIZATION--of H1DW1=webster--has H2DW2=itn--of ET12SameNP=PERSON--ORGANIZATION--f
alse ET12SamePP=PERSON--ORGANIZATION--false ET12SameVP=PERSON--ORGANIZATION--false orderM=1 HM1=Webster HM2=Itn HM12=Webster--Itn ET1=PERSON ET2=ORGANIZATION ET12=PERSON--ORGANIZATION EST12=Individual--Media ML1=NAM ML2=NAM ML12
=NAM--NAM NUMMB=0 ET12M1inM2=PERSON--ORGANIZATION--false ET12M2inM1=PERSON--ORGANIZATION--false HM12M1inM2=Webster--Itn--false HM12M2inM1=Webster--Itn--false detectorLabel=1 classLabel=ORG-AFF--Employment
BagWM1_ali=TRUE BagWM2_hospital=TRUE WBFL=and BM1F=in BM1L=weeks AM2F=is AM2L=recovering NUMWB=1 TPatternET=FACILITY_and_PERSON PTP=NP PTPH=NP--weeks CPHBM1F=in CPHBM1L=weeks CPHBNULL=true CPP=NP CPPH=NP--ali CPHAM2F=recovering
CPHAM2L=fast DPathET=FACILITY_conj_and_PERSON ET1DW1=PERSON--in ET2DW2=FACILITY--ali H1DW1=ali--in H2DW2=hospital--ali ET12SameNP=PERSON--FACILITY--true ET12SamePP=PERSON--FACILITY--true ET12SameVP=PERSON--FACILITY--false orderM
=2 HM1=Ali HM2=hospital HM12=Ali--hospital ET1=PERSON ET2=FACILITY ET12=PERSON--FACILITY EST12=Individual--Building-Grounds ML1=NAM ML2=NOM ML12=NAM--NOM NUMMB=0 ET12M1inM2=PERSON--FACILITY--false ET12M2inM1=PERSON--FACILITY--fa
lse HM12M1inM2=Ali--hospital--false HM12M2inM1=Ali--hospital--false detectorLabel=1 classLabel=PHYS--Located
```

Features for RE: Brown Word Clustering

- The Brown algorithm (a hierarchical clustering algorithm):
  - initially assigns each word to its own cluster
  - repeatedly merges the two clusters which cause the least loss in average mutual information between adjacent clusters based on bigram statistics
  - by tracing the pairwise merging steps, one can obtain a word hierarchy which can be represented as a binary tree
- Use prefixes of the bit strings of the heads of the entity mentions as the features (i.e., HM1_WC2, HM2_WC4)

Type	P		R		F	
	Baseline	PC4 (△)	Baseline	PC4 (△)	Baseline	PC4 (△)
EMP-ORG	75.4	77.2(+1.8)	79.8	81.5(+1.7)	77.6	79.3(+1.7)
PHYS	73.2	71.2(-2.0)	61.6	60.2(-1.4)	66.9	65.3(-1.7)
GPE-AFF	67.1	69.0(+1.9)	60.0	63.2(+3.2)	63.3	65.9(+2.6)
PER-SOC	88.2	83.9(-4.3)	58.4	61.0(+2.6)	70.3	70.7(+0.4)
DISC	79.4	80.6(+1.2)	42.9	46.0(+3.2)	55.7	58.6(+2.9)
ART	87.9	96.9(+9.0)	63.0	67.4(+4.4)	73.4	79.3(+5.9)
OTHER-AFF	70.6	80.0(+9.4)	41.4	41.4(0.0)	52.2	54.6(+2.4)

Bit string	Examples
111011011100	***US*** ...
1110110111011	***U.S.*** ...
1110110110000	***American*** ...
1110110111110110	***Cuban***, *Pakistani, Russian* ...
11111110010111	***Germany***, *Poland, Greece* ...
110111110100	***businessman***, *journalist, reporter*
1101111101111	***president,*** *governor, premier*...
1101111101100	***senator***, *soldier, ambassador* ...
11011101110	***spokesman***, *spokeswoman*, ...
11001100	***people***, *persons, miners, Haitians*
110110111011111	***base***, *compound, camps, camp* ...
110010111	***helicopters***, *tanks, Marines* ...

Sun at al., 2011: Semi-supervised Relation Extraction with Large-scale Word Clustering (ACL)

Features for RE: Word Embeddings

- Generalizing the head words of the entity mentions seems to be very helpful for RE
- Use word embeddings to achieve such generalization (i.e., using the dimensions of the word embeddings of the heads as the features)
- Without regularization:

System	In-domain	bc	cts	wl
Baseline(B)	51.4	49.7	41.5	36.6
B+WC10	52.3(+0.9)	50.8(+1.1)	45.7(+4.2)	39.6(+3)
B+WC	53.7(+2.3)	52.8(+3.1)	46.8(+5.3)	41.7(+5.1)
B+ED	54.1(+2.7)	52.4(+2.7)	46.2(+4.7)	42.5(+5.9)
B+WC+ED	**55.5(+4.1)**	**53.8(+4.1)**	**47.4(+5.9)**	**44.7(+8.1)**

- With regularization:

System	In-domain	bc	cts	wl
Baseline(B)	56.2	55.5	48.7	42.2
B+WC10	57.5(+1.3)	57.3(+1.8)	52.3(+3.6)	45.0(+2.8)
B+WC	58.9(+2.7)	58.4(+2.9)	52.8(+4.1)	47.3(+5.1)
B+ED	58.9(+2.7)	59.5(+4.0)	52.6(+3.9)	48.6(+6.4)
B+WC+ED	**59.4(+3.2)**	**59.8(+4.3)**	**52.9(+4.2)**	**49.7(+7.5)**

Kernel-based methods for RE

- Goal is to find training examples similar to test case
  - Need similarity metrics between pairs of relation instances
  - Determining similarity through features is awkward
  - Better to define a similarity measure directly: a kernel function
- Kernels can be used directly by
  - SVMs
  - Memory-based learners (k-nearest-neighbor)
- For RE, kernels defined over
  - Strings
  - Parse or Dependency Trees

String kernels

- Two strings are more similar if they share more substrings

Linear combination parameter

$$k(s_i, s_j) = \sum_n c_n k_n(s_i, s_j)$$

Decaying factor
$0 < \lambda \leq 1$

$$k_n(s_i, s_j) = \sum_{u \in \Sigma^n} \sum_{u = \mathbf{o}_{s_i}} \sum_{u = \mathbf{p}_{s_j}} \lambda^{l(\mathbf{o}_{s_i})} \lambda^{l(\mathbf{p}_{s_j})}$$

Sets of strings of length *n*

Length of the string

- Many variants are possible

Tree kernels

- Compute the number of common subtrees:

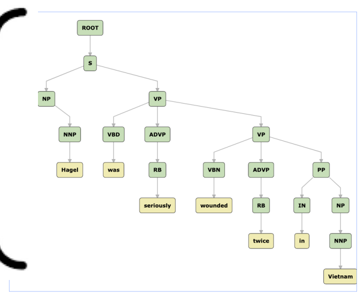
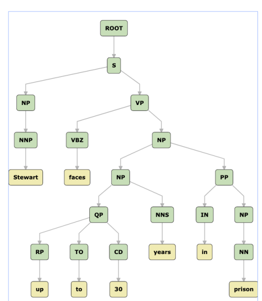
let N_1 and N_2 be the set of nodes in T_1 and T_2 respectively, then

$$TK_\sigma(T_1, T_2) = \sum_{n_1 \in N_1, n_2 \in N_2} \Delta(n_1, n_2)$$

where $\Delta(n_1, n_2)$ is computed by:

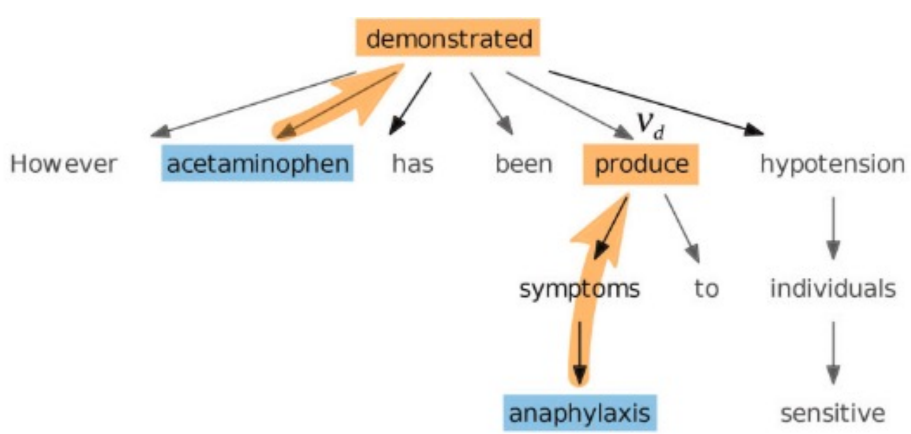
i) if n_1 and n_2 have different productions: $\Delta(n_1, n_2) = 0$; else
ii) if n_1 and n_2 are pre-terminals: $\Delta(n_1, n_2) = \lambda$; else
iii) $\Delta(n_1, n_2) = \lambda \prod_{j=1}^{nc(n_1)} (1 + \Delta(ch(n_1, j), ch(n_2, j)))$

- T_1, T_2 can be either constituent or dependency trees. The trees can be pruned to minimally cover the two entity mention of interest.
- Can incorporate with word clusters and word embeddings

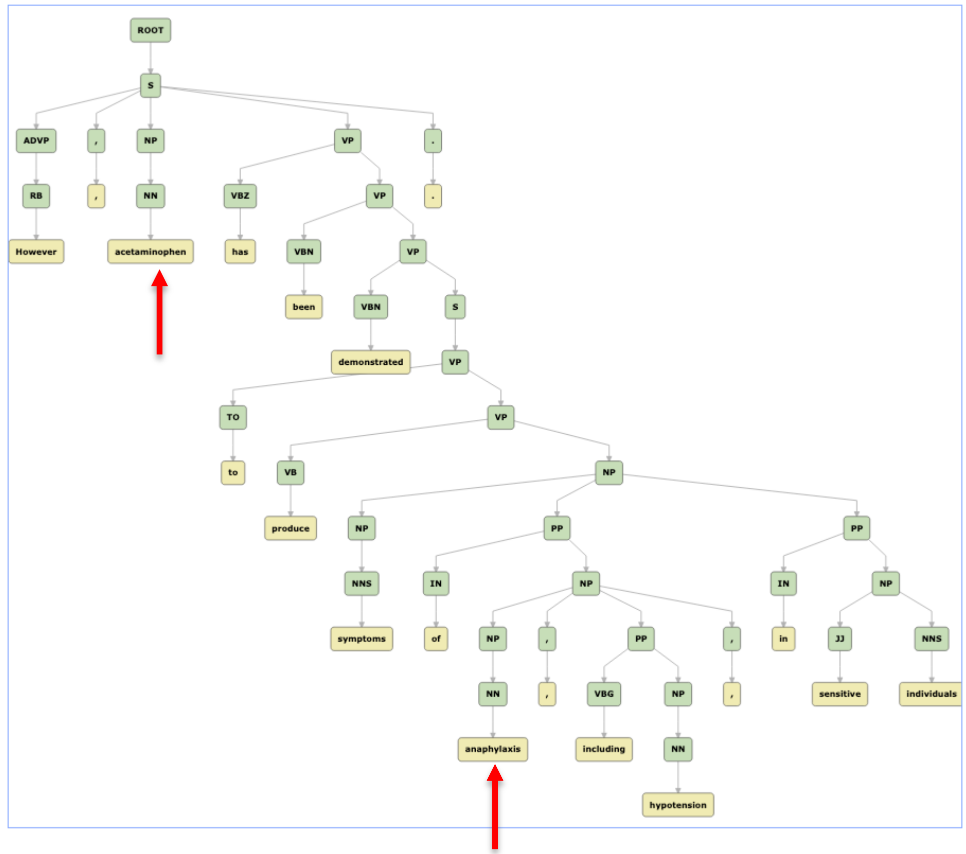
$\Phi\{$  ,  $\} = 0.95$

Tree kernels

However, acetaminophen has been demonstrated to produce symptoms of anaphylaxis, including hypotension, in sensitive individuals.



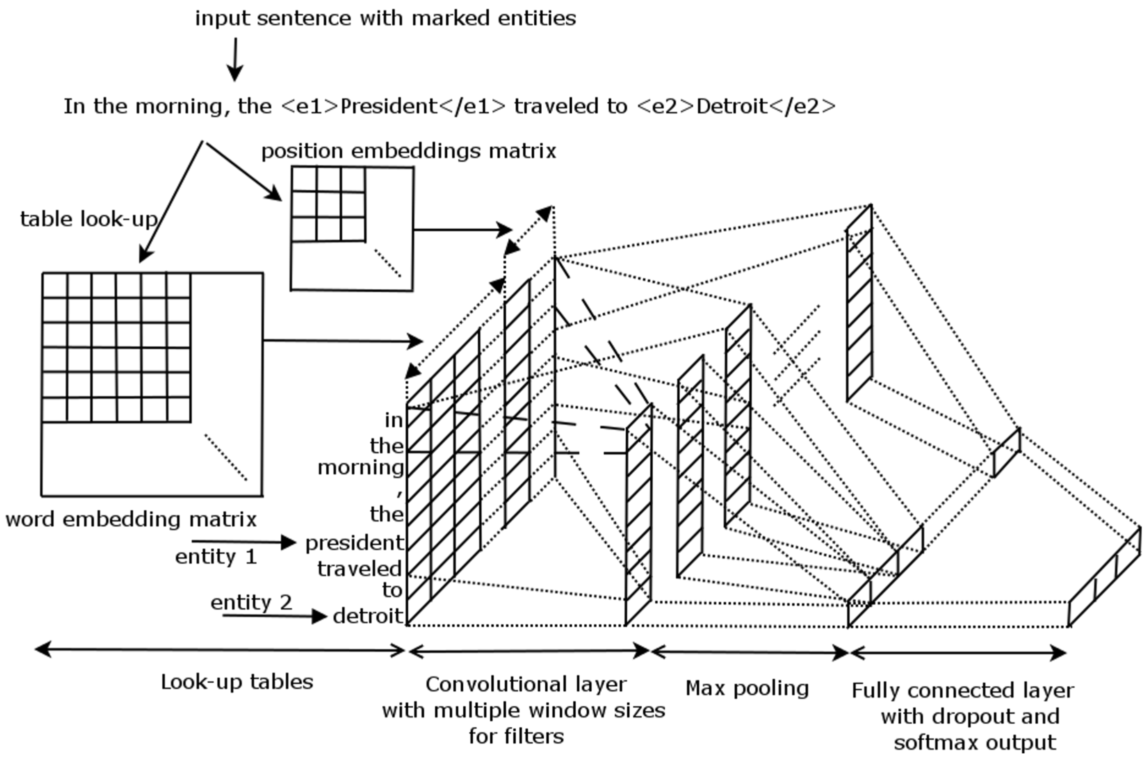
The dependency tree



The constituent tree

Deep learning for RE

- Avoid feature or kernel design for RE



A Convolutional Neural Network (CNN) for Relation Extraction

Classifier	Features	F
MaxEnt	POS, WordNet, morphological features, noun compound sys- tem, thesauri, Google n -grams	77.6
SVM	POS, WordNet, prefixes and other morphological fea- tures, dependency parse, Levin classes, PropBank, FrameNet, NomLex-Plus, Google n -grams, paraphrases, TextRunner	82.2
CNN (Zeng et al., 2014)	WordNet	82.7
CNN (Nguyen and Grishman, 2015a)	-	82.8

SemEval 2010 Dataset

Nguyen and Grishman, 2015

Position embeddings

- To inform the models about the two entity mentions of interest, we introduce (relative) position embeddings (randomly initialized and updated during training)

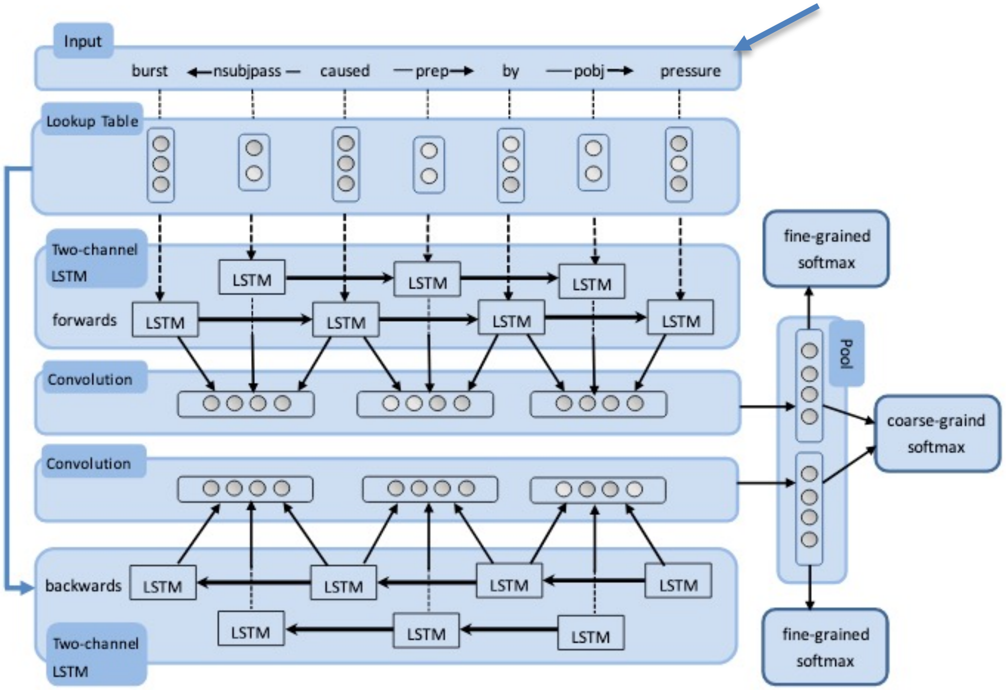
dist from m1	0	1	3	4	5	6	7	8	9
dist from m2	-8	-7	-6	-5	-4	-3	-2	-1	0
	[The Big Sleep]	is	a	1946	film	noir	directed	by	[Howard Hawks]

-4	2	-0.5	1.1	0.3	0.4	-0.5
-3	-1.4	0.4	-0.2	-0.9	0.5	0.9
-2	-1.1	-0.2	-0.5	0.2	-0.8	0
-1	0.7	-0.3	1.5	-0.3	-0.4	0.1
0	-0.8	1.2	1	-0.7	-1	-0.4
1	0	0.3	-0.3	-0.9	0.2	1.4
2	0.8	0.8	-0.4	-1.4	1.2	-0.9
3	1.6	0.4	-1.1	0.7	0.1	1.6
4	1.2	-0.2	1.3	-0.4	0.3	-1.0

Deep learning for RE

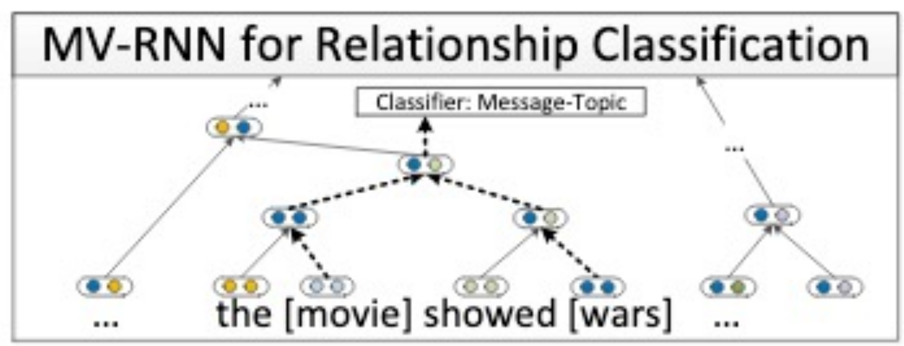
- Can also incorporate syntax into deep learning models for RE: to identify important context words (i.e., via the dependency paths) or to guide the computational flows of the neural network models.

the shortest dependency path between two entity mentions



Recursive neural networks:
building the networks based on
the constituent trees

the binarized constituent subtree

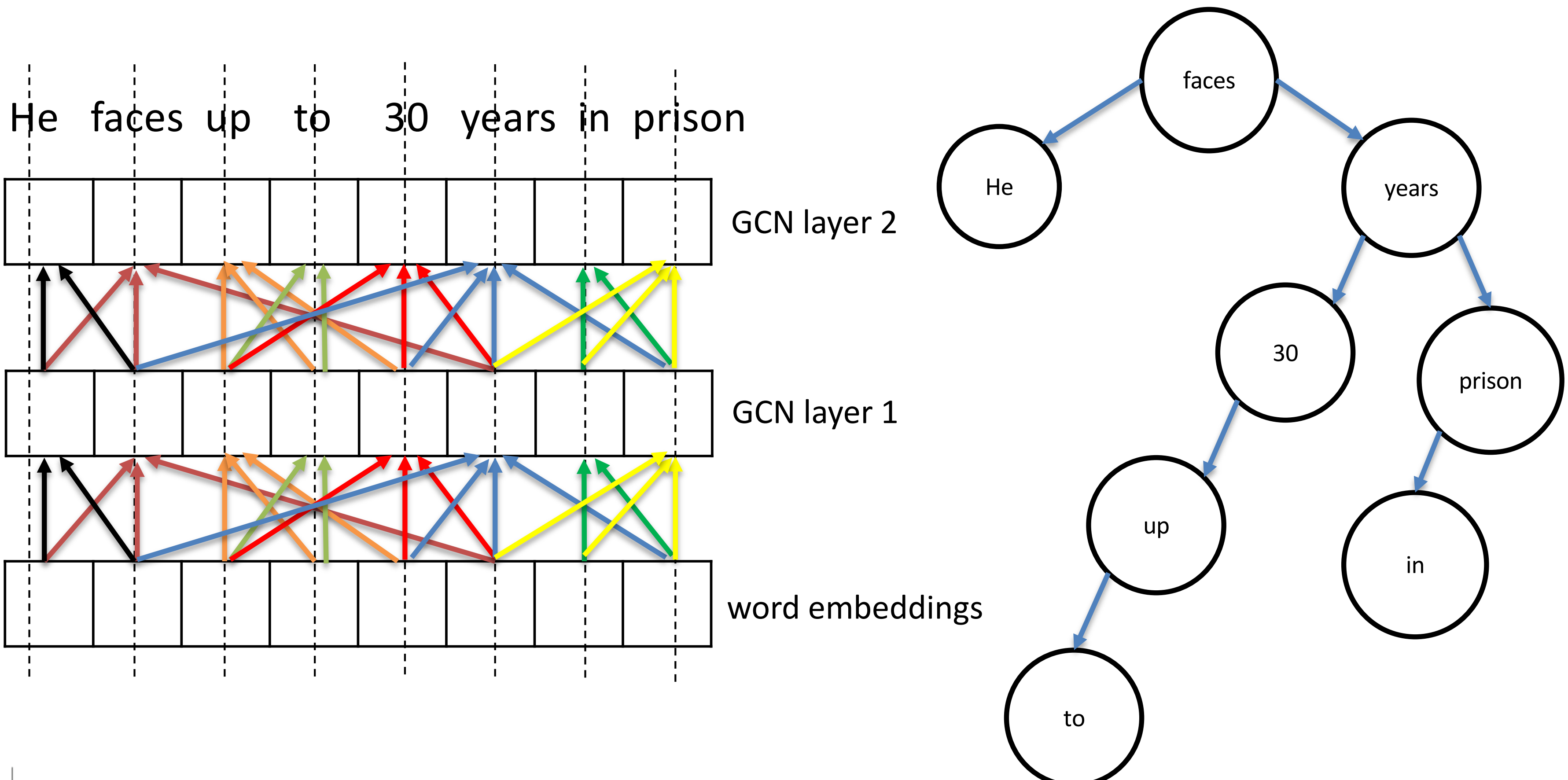


Cat et al., Bidirectional Recurrent Convolutional Neural Network for Relation Classification (ACL 2016)

Socher et al., Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank (EMNLP 2013)

Syntactic Structures for Relation Extraction

- Graph Convolutional Neural Network (GCN) over dependency trees for RE (a recent state-of-the-art approach for RE) (Zhang et al., 2018)



Other types of learning for IE

- Supervised learning
  - All training data is labeled

- Semi-supervised learning
  - Part of training data is labeled ('the seed') (the rest is unlabeled)
  - Make use of redundancies to learn labels of additional data, then train model
  - Co-training
  - Reduces amount of data which must be hand-labeled to achieve a given level of performance

- Active learning
  - Start with partially labeled data
  - System selects additional 'informative' examples for users to label
  - Information examples can be selected via uncertainty scores, committee disagreement, representativeness (i.e., frequency of features), or diversity.

Semi-supervised learning

L = labeled data

U = unlabeled data

1. L = seed
   -- repeat 2-4 until stopping condition is reached
2. C = classifier trained on L
3. Apply C to U.
   N = most confidently labeled items
4. L += N; U -= N

Confidence

How to estimate confidence?

- Binary probabilistic classifier
  - Confidence = | $P - 0.5$ | * 2
- N-ary probabilistic classifier
  - Confidence = $P_1 - P_2$

  where

  P_1 = probability of most probable label

  P_2 = probability of second most probable label
- SVM
  - Distance from the separating hyperplane

Co-training

- Two 'views' of data (subsets of features)
  - Producing two classifiers $C_1(x)$ and $C_2(x)$
- Ideally
  - Independent
  - Each sufficient to classify data
- Apply classifiers in alternation (or in parallel)

1. L = seed
   -- repeat 2-7 until stopping condition is reached
2. C_1= classifier trained on L
3. Apply C_1 to U.
   N = most confidently labeled items
4. L += N; U -= N
5. C_2 = classifier trained on L
6. Apply C_2 to U.
   N = most confidently labeled items
7. L += N; U -= N

Problems with semi-supervised learning

- When to stop?
  - U is exhausted
  - Reach performance goal using held-out labeled sample
  - After fixed number of iterations based on similar tasks
- Poor confidence estimates
  - Errors from poorly-chosen data rapidly magnified

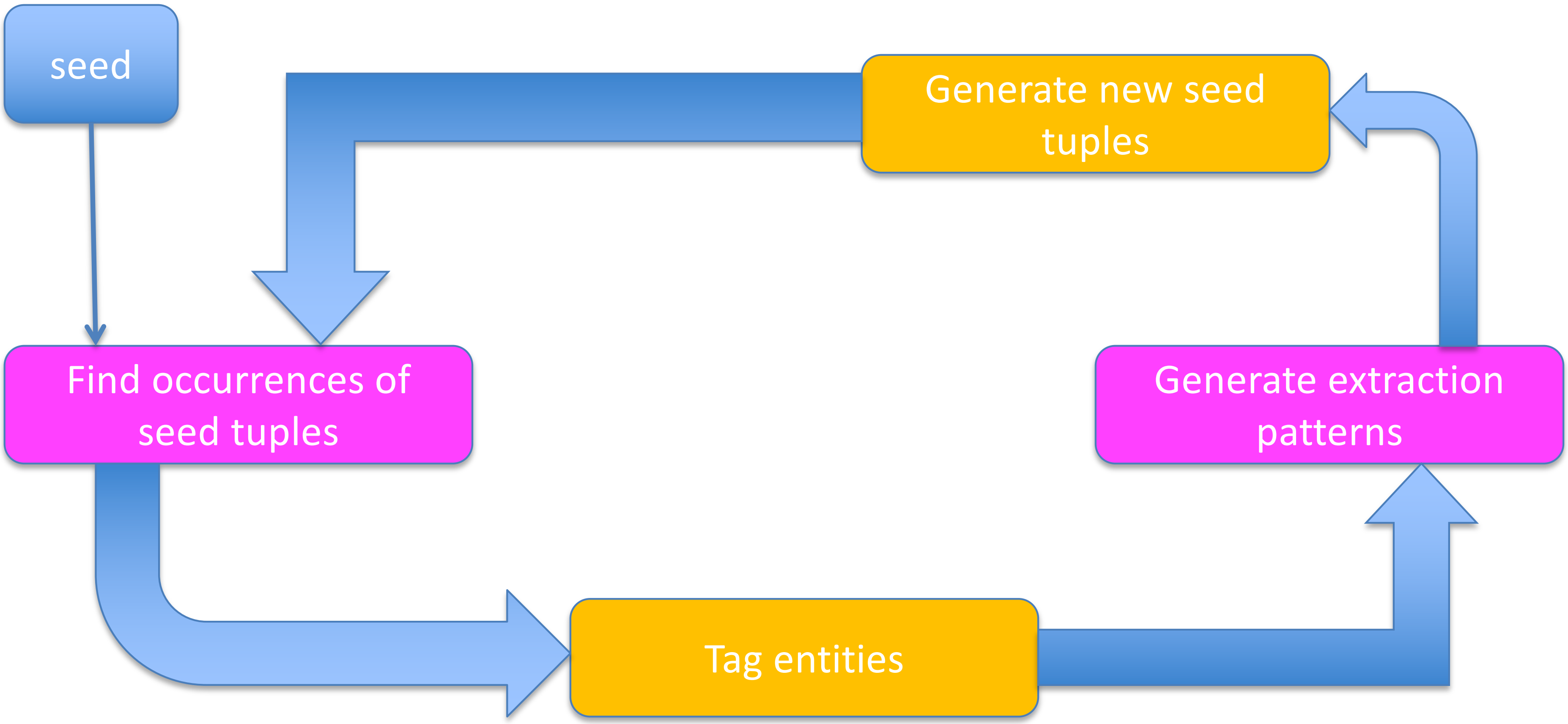
Semi-supervised methods for RE

- Preparing training data for relations is more costly than for names
  - Must annotate entities and relations
- So there is a strong motivation to minimize training data through semi-supervised methods
- As for names, we will adopt a co-training approach:
  - Feature set 1: the two entities
  - Feature set 2: the contexts between the entities
- We will limit the bootstrapping
  - to a specific pair of entity types
  - and to instances where both entities are named

Semi-supervised learning for RE

- Seed:
  - [*Moby Dick,* Herman Melville]
- Contexts for seed:
  - … wrote …
  - … is the author of …
- Other pairs appearing in these contexts
  - [*Animal Farm*, George Orwell]
  - [*Don Quixote,* Miguel de Cervantes]
- Additional contexts …

Co-training for relations



Ranking contexts

- If relation R is functional,
  and $[X, Y]$ is a seed,
  then $[X, Y']$, $Y' \neq Y$, is a negative example

- Confidence of pattern P

$$Conf(P) = \frac{P.positive}{P.positive + P.negative}$$

- Where

$P.positive$ = number of positive matches to pattern P

$P.negative$ = number of negative matches to pattern P

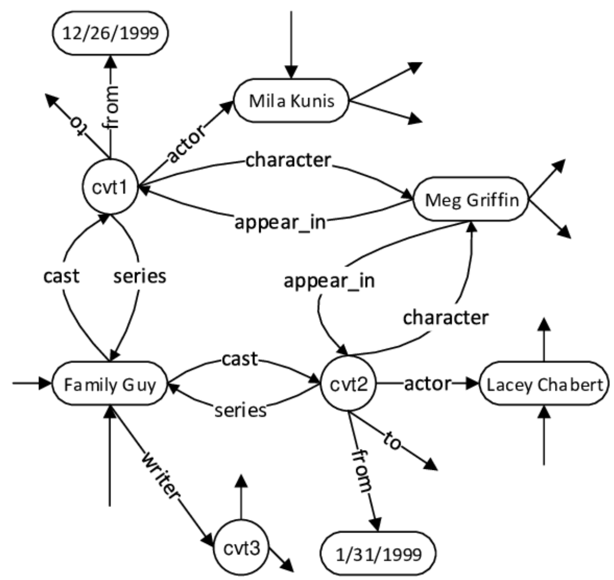
Ranking pairs

- Once a confidence has been assigned to each pattern, we can assign a confidence to each new pair based on the patterns in which it appears
  - Confidence of best pattern
  - Combination assuming patterns are independent

$$Conf(X,Y) = 1 - \prod_{P \in contexts\_of\_(X,Y)} (1 - Conf(P))$$

Distant supervision

- Sometimes a large database is available involving the type of relations to be extracted
    - A number of such public data bases are now available, such as FreeBase and Wiki Infobox
- Text instances corresponding to some of the database instances can be found in a large corpus or from the Web
- Together these can be used to train a relation classifier



Distant supervision

Ronaldinho

From Wikipedia, the free encyclopedia

"Ronaldinho Gaucho" redirects here. For the comic strip based on him, see Ronaldinho Gaucho (comic strip). For other uses, see Ronaldinho (disambiguation).

This name uses Portuguese naming customs: the first or maternal family name is Assis *and the second or paternal family name is* Moreira.

Ronaldo de Assis Moreira (born 21 March 1980), commonly known as **Ronaldinho Gaúcho** (Brazilian Portuguese: [ʁonawˈdʒĩɲu ɡaˈuʃu]) or simply **Ronaldinho**,[note 1] is a Brazilian former professional footballer and ambassador for Barcelona.[4] He played mostly as an attacking midfielder, but was also deployed as a forward or a winger. He played the bulk of his career at European clubs Paris Saint-Germain, Barcelona and A.C. Milan as well as playing for the Brazilian national team. Often considered one of the best players of his generation and regarded by many as one of the greatest of all time,[note 2] Ronaldinho won two FIFA World Player of the Year awards and a Ballon d'Or. He was renowned for his technical skills and creativity; due to his agility, pace and dribbling ability, as well as his use of tricks, feints, overhead kicks, no-look passes and accuracy from free-kicks.

Ronaldinho made his career debut for Grêmio, in 1998. At age 20, he moved to Paris Saint-Germain in France before signing for Barcelona in 2003. In his second season with Barcelona, he won his first FIFA World Player of the Year award, as Barcelona won La Liga. The season that followed is considered one of the best in his career as he was instrumental in Barcelona winning the UEFA Champions League, their first in fourteen years, as well as another La Liga title, giving Ronaldinho his first career double. After scoring two spectacular solo goals in *El Clásico*, Ronaldinho became the second Barcelona player, after Diego Maradona in 1983, to receive a standing ovation from Real Madrid fans at the Santiago Bernabéu. Ronaldinho also received his second FIFA World Player of the Year award, as well as the Ballon d'Or.

Ronaldinho

Ronaldinho in 2019

Personal information		
Full name	Ronaldo de Assis Moreira[1]	
Date of birth	21 March 1980 (age 39)[1]	
Place of birth	Porto Alegre, Brazil	
Height	1.81 m (5 ft 11 in)[1]	
Playing position	Attacking midfielder / Forward	
Youth career		
1987–1998	Grêmio	
Senior career*		
Years	**Team**	**Apps (Gls)**
1998–2001	Grêmio	52 (21)
2001–2003	Paris Saint-Germain	55 (17)
2003–2008	Barcelona	145 (70)
2008–2011	A.C. Milan	76 (20)
2011–2012	Flamengo	33 (15)
2012–2014	Atlético Mineiro	48 (16)
2014–2015	Querétaro	25 (8)
2015	Fluminense	7 (0)

Distant supervision: approach

- Given:
    - Database for relation R
    - Corpus containing information about relation R
- Collect $< X, Y >$ pairs from data base relation R
- Collect sentences in corpus containing both X and Y
    - These are positive training examples
- Collect sentences in corpus containing X and some Y' with the same entity type as Y such that $< X, Y' >$ is not in the data base
    - These are negative training examples
- Use examples to train classifier which operates on pairs of entities

Freebase

Relation	Entity1	Entity2
/business/company/founders	Apple	Steve Jobs
...	...	...

Mentions from free texts

1. Steve Jobs was the co-founder and CEO of Apple and formerly Pixar.
2. Steve Jobs passed away the day before Apple unveiled iPhone 4S in late 2011.

Distant supervision: limitations

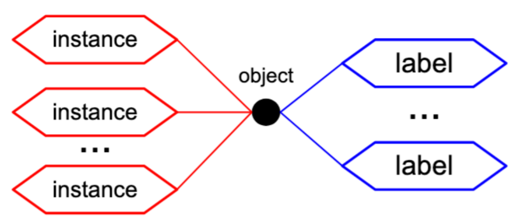
- The training data produced through distant supervision may be quite noisy:
  - Given a pair $< X, Y >$ that is involved in multiple relations, $R < X, Y >$ and $R' < X, Y >$. If the database only caputure relation R and the text instance actually represents relation R', it will yield a false positive training instance
  - If many $< X, Y >$ pairs are involved, the classifier may learn the wrong relation
  - If a relation is incomplete in the data base … for example, if *resides_in* $< X, Y >$ contains only a few of the locations where a person has resided … then we will generate many false negatives, possibly leading the classifier to learn no relation at all

Multi-label multi-instance learning for distant supervision (MIML)

- To reduce noise in distant supervision:
  - Group instances (sentences) corresponding to the same entity pair $< X, Y >$ in the knowledge base into the a group (a bag of instances)
  - Each bag can be assigned to multiple relations to capture the possible relations between X and Y in the knowledge base.
  - People might just do multi-instance learning (i.e., a single label for a bag)

$$\mathrm{DB} = \begin{pmatrix} \mathit{BornIn}(\text{Barack Obama, United States}) \\ \mathit{EmployedBy}(\text{Barack Obama, United States}) \end{pmatrix}$$

Sentence	Latent Label
Barack Obama is the 44th and current President of the United States.	*EmployedBy*
Obama was born in the United States just as he has always said.	*BornIn*
United States President Barack Obama meets with Chinese Vice President Xi Jinping today.	*EmployedBy*
Obama ran for the United States Senate in 2004.	–



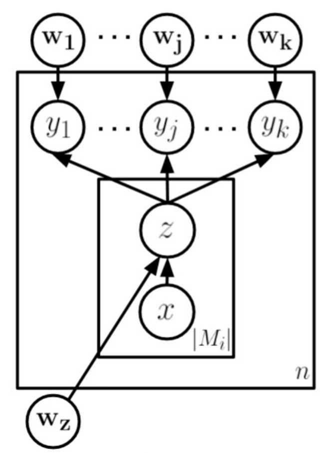
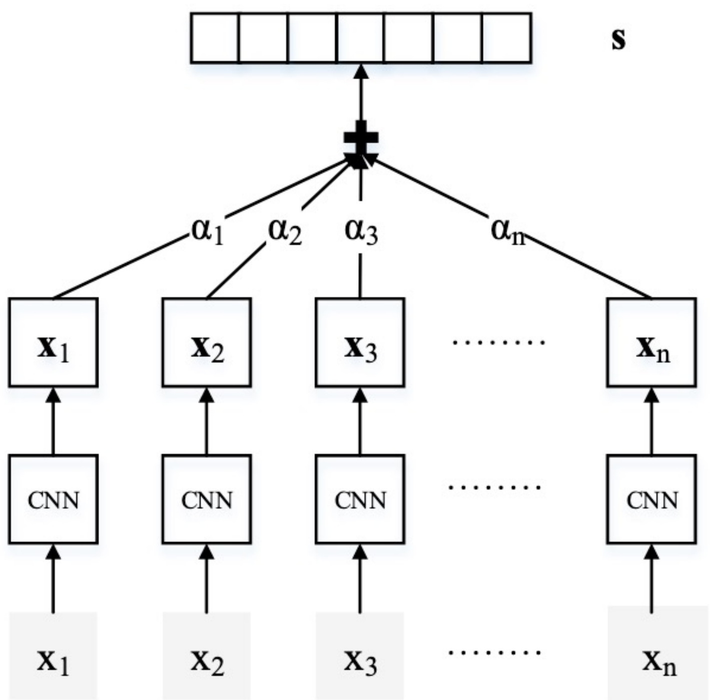


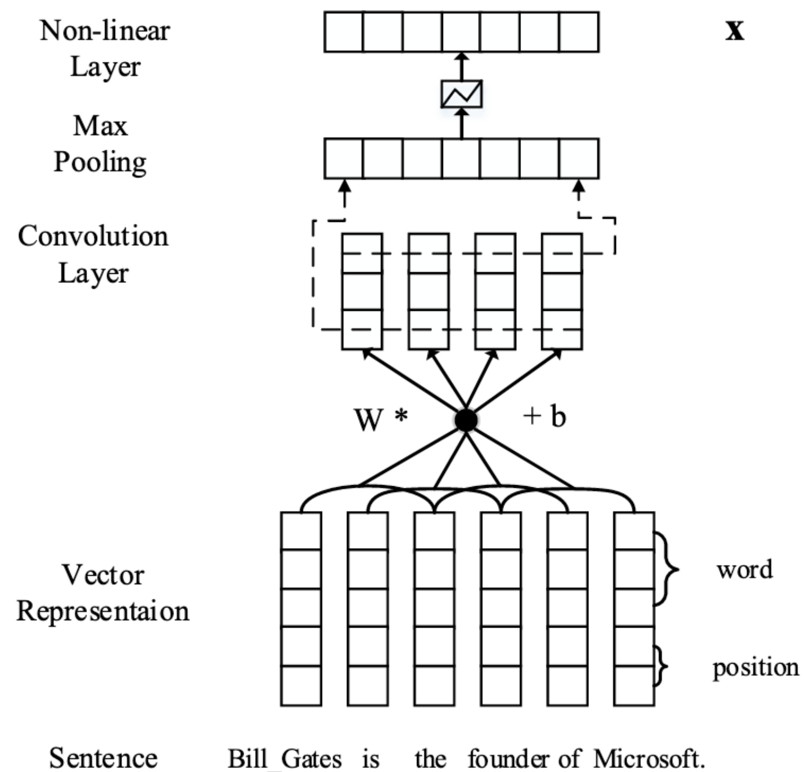
Figure 3: MIML model plate diagram. We unrolled the y plate to emphasize that it is a collection of binary classifiers (one per relation label), whereas the z classifier is multi-class. Each z and y_j classifier has an additional prior parameter, which is omitted here for clarity.

Surdeanu et al., Multi-instance Multi-label Learning for Relation Extraction (EMNLP 2012)

Multiple-instance learning for distant supervision

$$\mathbf{s} = \sum_i \alpha_i \mathbf{x}_i \qquad \alpha_i = \frac{\exp(e_i)}{\sum_k \exp(e_k)} \qquad e_i = \mathbf{x}_i \mathbf{A} \mathbf{r} \qquad \mathbf{o} = \mathbf{M}\mathbf{s} + \mathbf{d} \qquad p(r|S,\theta) = \frac{\exp(o_r)}{\sum_{k=1}^{n_r} \exp(o_k)}$$





Lin et al., Neural Relation Extraction with Selective Attention over Instances (ACL 2016)