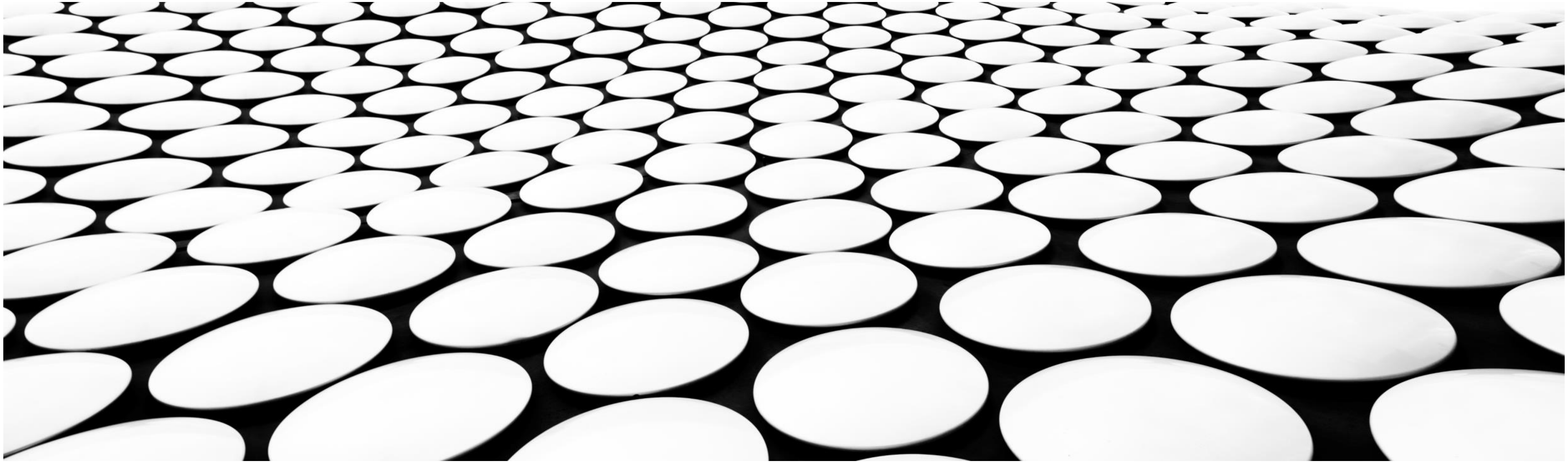
DISTANT SUPERVISION FOR RELATION EXTRACTION WITHOUT LABELED DATA (2009)

MIKE MINTZ, STEVEN BILLS, RION SNOW, DAN JURAFSKY

RODERICK PEREZ, PH.D

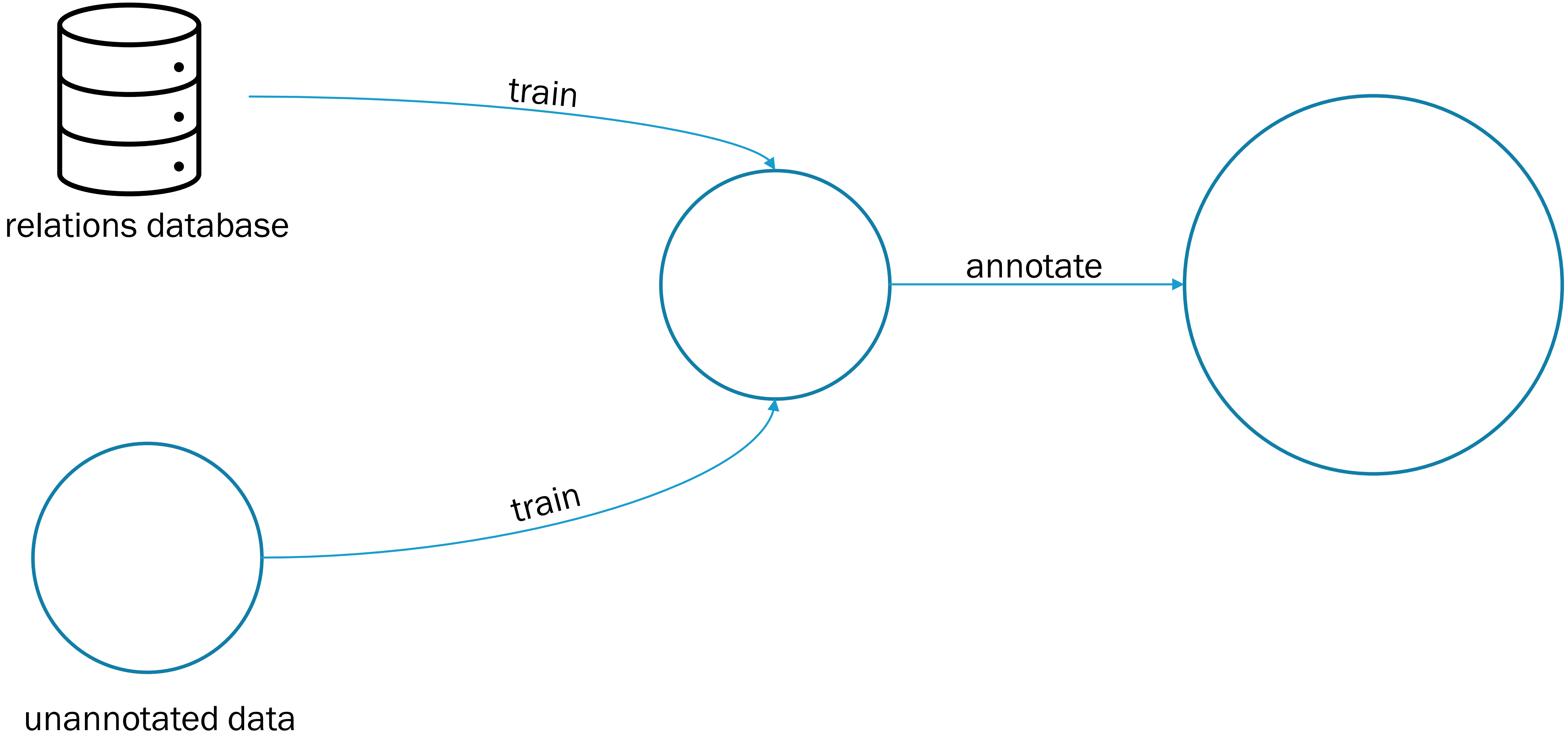
APRIL, 22TH, 2021



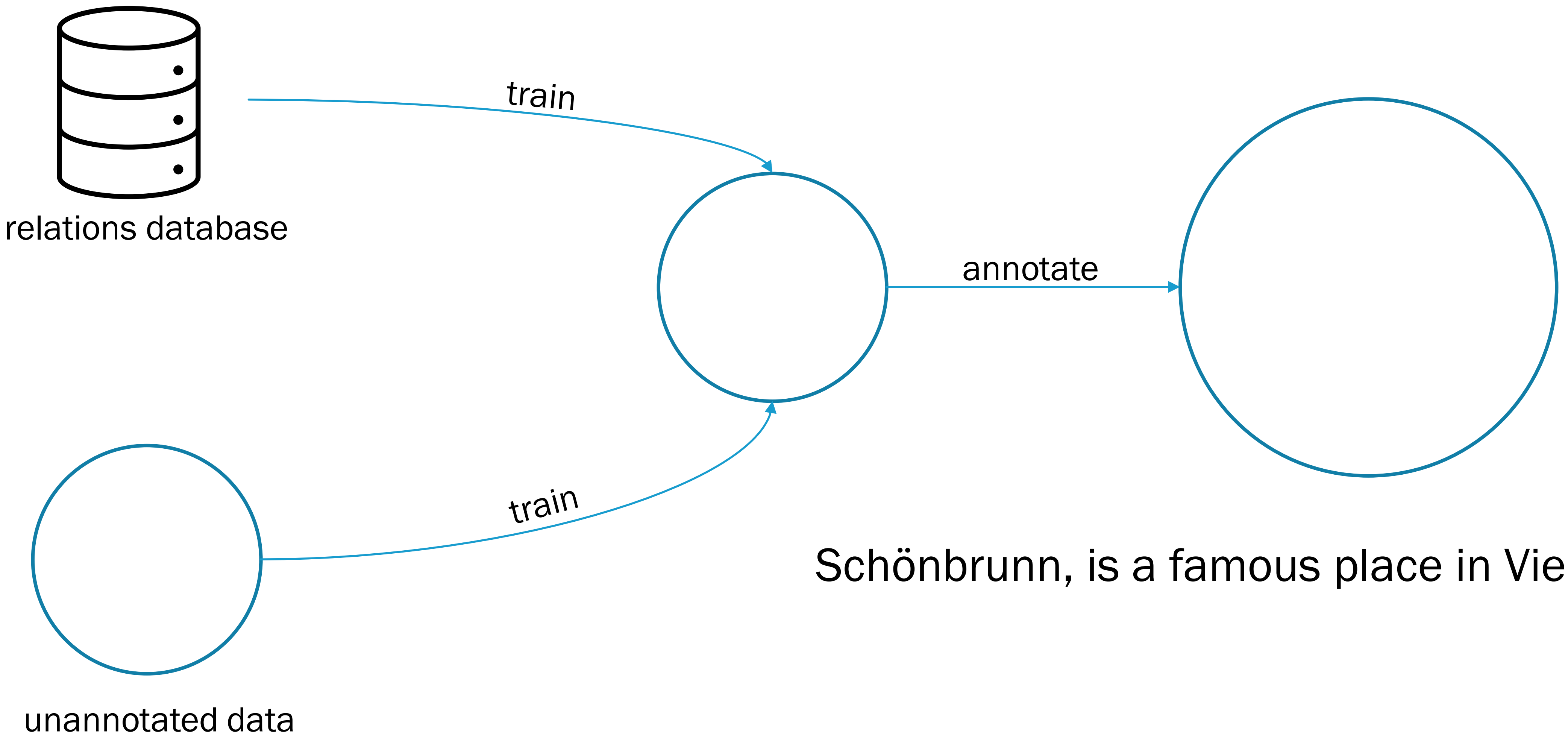
OUTLINE

- Summary
- Introduction
  - Entities, relations and events
- Motivation
- Relation Extraction
- FreeBase
- Approaches
- Training and Testing
- Results
- Conclusions
- Discussion

DISTANT SUPERVISION
SUMMARY

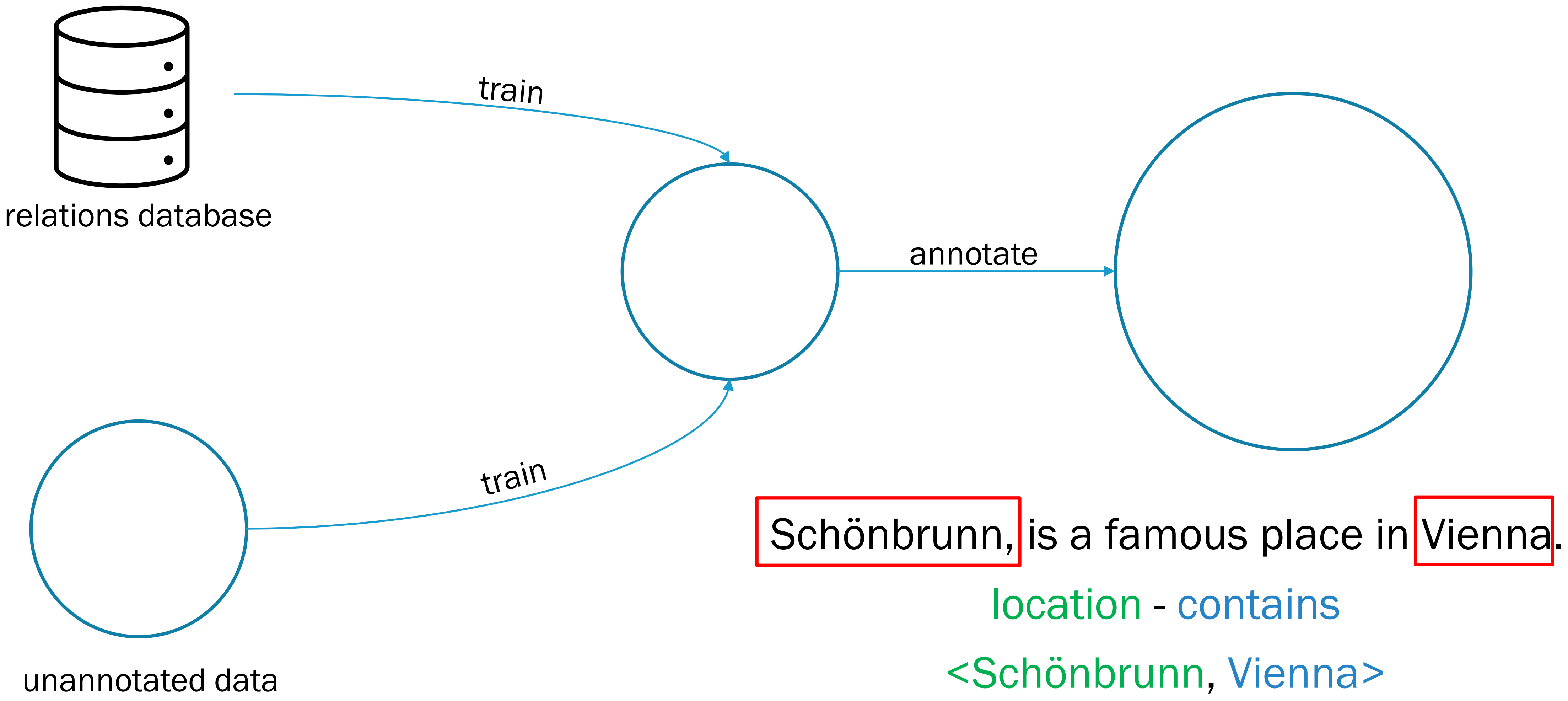


DISTANT SUPERVISION
SUMMARY



Schönbrunn, is a famous place in Vienna.

DISTANT SUPERVISION
SUMMARY



ENTITIES, RELATIONS AND EVENTS

Entities: persons, organizations, locations, facilities, weapons, vehicles, and geo-political entities.

38 years old Roderick Perez study a Master in Data Science at the University of Vienna is the father of Lea and Amy...

ENTITIES, RELATIONS AND EVENTS

Entities: persons, organizations, locations, facilities, weapons, vehicles, and geo-political entities.

Relations (between entities): an ordered, binary relation between entities. For example: role, part, located, near, and social.

38 years old Roderick Perez study a Master in Data Science at the University of Vienna is the $\boxed{\text{father of}}$ Lea and Amy...

X is the **father of** Y and Z

ENTITIES, RELATIONS AND EVENTS

Entities: persons, organizations, locations, facilities, weapons, vehicles, and geo-political entities.

Relations (between entities): an ordered, binary relation between entities. For example: role, part, located, near, and social.

Events: interaction, movement, transfer, creation and destruction.

38 years old Roderick Perez study a Master in Data Science at

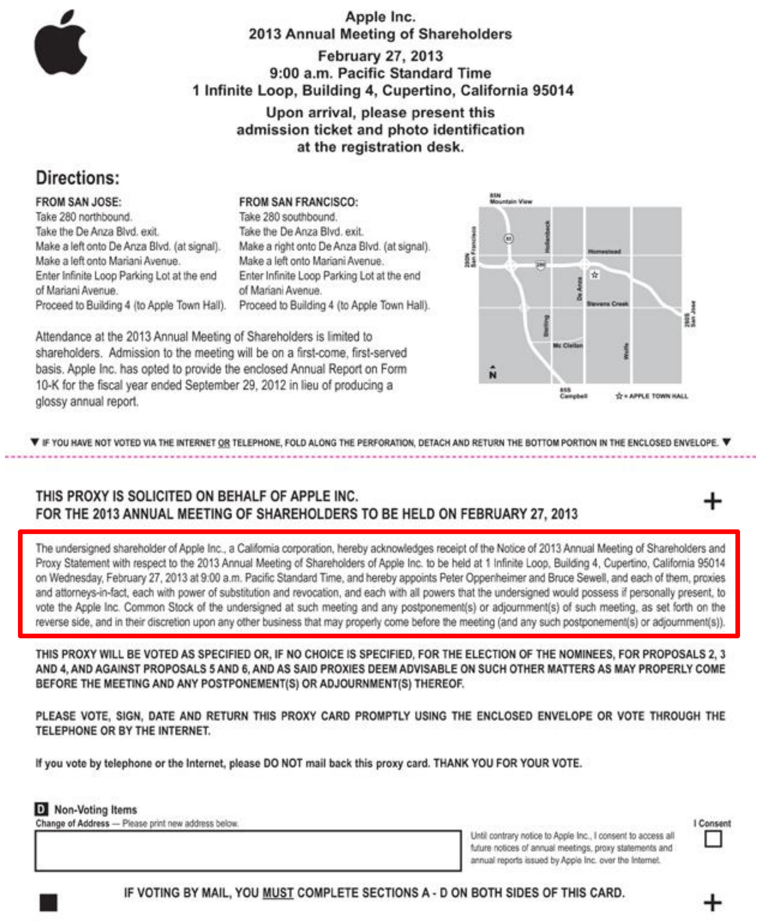
the University of Vienna is the father of Lea and Amy…

X **study** at Y

MOTIVATION

WHAT IS RELATION EXTRACTION?

Company Report Meeting [2013]



Apple Inc.
2013 Annual Meeting of Shareholders
February 27, 2013
9:00 a.m. Pacific Standard Time
1 Infinite Loop, Building 4, Cupertino, California 95014

Upon arrival, please present this admission ticket and photo identification at the registration desk.

Directions:

FROM SAN JOSE:
Take 280 northbound.
Take the De Anza Blvd. exit.
Make a left onto De Anza Blvd. (at signal).
Make a left onto Mariani Avenue.
Enter Infinite Loop Parking Lot at the end of Mariani Avenue.
Proceed to Building 4 (to Apple Town Hall).

FROM SAN FRANCISCO:
Take 280 southbound.
Take the De Anza Blvd. exit.
Make a right onto De Anza Blvd. (at signal).
Make a left onto Mariani Avenue.
Enter Infinite Loop Parking Lot at the end of Mariani Avenue.
Proceed to Building 4 (to Apple Town Hall).

Attendance at the 2013 Annual Meeting of Shareholders is limited to shareholders. Admission to the meeting will be on a first-come, first-served basis. Apple Inc. has opted to provide the enclosed Annual Report on Form 10-K for the fiscal year ended September 29, 2012 in lieu of producing a glossy annual report.

▼ IF YOU HAVE NOT VOTED VIA THE INTERNET OR TELEPHONE, FOLD ALONG THE PERFORATION, DETACH AND RETURN THE BOTTOM PORTION IN THE ENCLOSED ENVELOPE. ▼

THIS PROXY IS SOLICITED ON BEHALF OF APPLE INC.
FOR THE 2013 ANNUAL MEETING OF SHAREHOLDERS TO BE HELD ON FEBRUARY 27, 2013

The undersigned shareholder of Apple Inc., a California corporation, hereby acknowledges receipt of the Notice of 2013 Annual Meeting of Shareholders and Proxy Statement with respect to the 2013 Annual Meeting of Shareholders of Apple Inc. to be held at 1 Infinite Loop, Building 4, Cupertino, California 95014 on Wednesday, February 27, 2013 at 9:00 a.m. Pacific Standard Time, and hereby appoints Peter Oppenheimer and Bruce Sewell, and each of them, proxies and attorneys-in-fact, each with power of substitution and revocation, and each with all powers that the undersigned would possess if personally present, to vote the Apple Inc. Common Stock of the undersigned at such meeting and any postponement(s) or adjournment(s) of such meeting, as set forth on the reverse side, and in their discretion upon any other business that may properly come before the meeting (and any such postponement(s) or adjournment(s)).

THIS PROXY WILL BE VOTED AS SPECIFIED OR, IF NO CHOICE IS SPECIFIED, FOR THE ELECTION OF THE NOMINEES, FOR PROPOSALS 2, 3 AND 4, AND AGAINST PROPOSALS 5 AND 6, AND AS SAID PROXIES DEEM ADVISABLE ON SUCH OTHER MATTERS AS MAY PROPERLY COME BEFORE THE MEETING AND ANY POSTPONEMENT(S) OR ADJOURNMENT(S) THEREOF.

PLEASE VOTE, SIGN, DATE AND RETURN THIS PROXY CARD PROMPTLY USING THE ENCLOSED ENVELOPE OR VOTE THROUGH THE TELEPHONE OR BY THE INTERNET.

If you vote by telephone or the Internet, please DO NOT mail back this proxy card. THANK YOU FOR YOUR VOTE.

D Non-Voting Items
Change of Address — Please print new address below.

Until contrary notice to Apple Inc., I consent to access all future notices of annual meetings, proxy statements and annual reports issued by Apple Inc. over the Internet. I Consent ☐

IF VOTING BY MAIL, YOU MUST COMPLETE SECTIONS A - D ON BOTH SIDES OF THIS CARD.

Source: SEC.gov

"The undersigned shareholder of Apple Inc., a California corporation, hereby acknowledges receipt of the Notice 2013 Annual Meeting of Shareholders and Proxy Statement with respect to 2013 Annual Meeting Shareholders of Apple Inc. to be held at Infinite Loop, Building 4, Cupertino, California 95014 on Wednesday, February 27, 2013 at 9:00a.m. Pacific Standard Time…"

Extracted Complex Relation:

- Company – Meeting Invitation
  - Company: Apple Inc.
  - Location: Infinite Loop, …, California 95014
  - Date: February 27, 2013
  - Time: 9:00a.m. Pacific Standard Time

Relation triples:

- meeting – location (Apple, California)
- meeting – year (Apple, 2013)
- meeting – time (Apple, 9:00a.m.)

ABSTRACT

Distant supervision for relation extraction without labeled data (2009)

Authors: Mike Mintz, Steven Bills, Rion Snow, Dan Jurafsky

- Alternative that does not require labeled corpora
- Avoid domain dependence of Automatic Content Extraction (ACE) -> Supervised Learning
- Allowing the use of corpora of any size

- Data Source:
  - FreeBase

- Results:
  - Extract 10,000 instances
  - Precision: 67.6%
  - Syntactic parse features are helpful for relations that are ambiguous or lexically distant

FREEBASE

entity1	verb	verb	preposition	entity2		entity3
Edwin Hubble	was	born	in	Marshfield	,	Missouri

Astronomer Edwin Hubble was born in Marshfield , Missouri

Freebase <u>only</u> contains **NODES** and **LINKS**

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

The 23 largest Freebase relations we use, with their size and an instance of each relation

[Source: Mintz et al., 2009, Distant supervision for relation extraction without labeled data

HOW TO BUILD RELATION EXTRACTORS?

APPROACHES

- Manual
- Supervised (ACE)
- Unsupervised

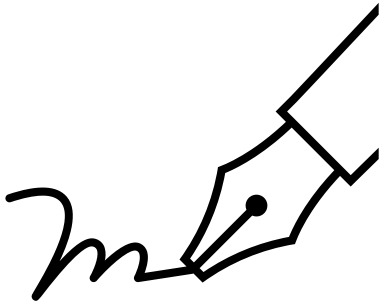
RELATION EXTRACTION

PER GPE
Roderick Perez was in Colombia → Person - Located

ORG ORG
Alphabet, the parent company of Google → Organization - Organization

PER ORG
Bill Gates, co-founder of Microsoft → Person - Organization

APPROACHES
MANUAL



PER GPE

Roderick Perez was in Colombia → Person - Located

Positive	Negative
▪ High-precision ▪ Tailored to specific domains	▪ Low-recall ▪ A lot of work for all possible patterns! ▪ Don’t want to have to do this for every relation! ▪ We’d like better accuracy

APPROACHES
SUPERVISED

Astronomer	Edwin Hubble	was	born	in	Marshfield	,	Missouri

APPROACHES

SUPERVISED

	entity1	verb	verb	preposition	entity2		entity3
Astronomer	Edwin Hubble	was	born	in	Marshfield	,	Missouri

1. Find all pairs of named entities

APPROACHES
SUPERVISED

	entity1	verb	verb	preposition	entity2		entity3
Astronomer	Edwin Hubble	was	born	in	Marshfield	,	Missouri

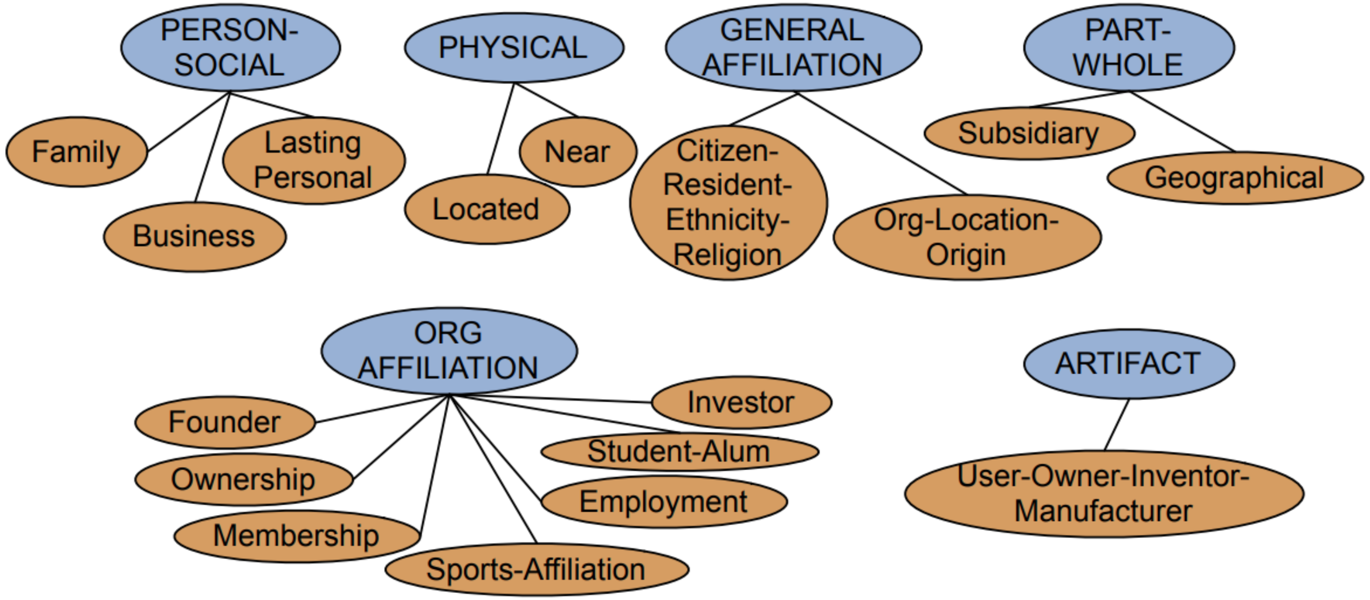
1. Find all pairs of named entities
2. Decide if 2 entities are related

APPROACHES

SUPERVISED

1. Find all pairs of named entities
2. Decide if 2 entities are related
3. If yes, classify the relation

Automated Content Extraction (ACE)



APPROACHES

SUPERVISED

	entity1	verb	verb	preposition	entity2		entity3
Astronomer	Edwin Hubble	was	born	in	Marshfield	,	Missouri

Positive	Negative
High accuracies (large hand-labeled training data)	Labeling is expensive Supervised models are brittle Use only annotated data to train a model.

APPROACHES

BOOTSTRAPPING

1. Start with 5 seeds:

Tuple: <author, book>

Author	Book
Isaac Asimov	The Robots of Dawn
David Grin	Startide Rising
James Gleick	Chaos: Making a New Science
Charles Dickems	Great Expectations
William Shakespeare	The Comedy of Errors

Brin, Sergei. 1998. Extracting Patterns and Relations from the World Wide Web

APPROACHES
BOOTSTRAPPING

1. Start with 5 seeds:

Tuple: <author, book>

Author	Book
Isaac Asimov	The Robots of Dawn
David Grin	Startide Rising
James Gleick	Chaos: Making a New Science
Charles Dickems	Great Expectations
William Shakespeare	The Comedy of Errors

Brin, Sergei. 1998. Extracting Patterns and Relations from the World Wide Web

2. Find Instances:

- The Comedy of Error, by William Shakespeare, was…
- The Comedy of Error, by William Shakespeare, is…
- The Comedy of Error, one of William Shakespeare's earliest attempts…
- The Comedy of Error, one of William Shakespeare's most…

APPROACHES

BOOTSTRAPPING

1. Start with 5 seeds:

Tuple: <author, book>

Author	Book
Isaac Asimov	The Robots of Dawn
David Grin	Startide Rising
James Gleick	Chaos: Making a New Science
Charles Dickems	Great Expectations
William Shakespeare	The Comedy of Errors

Brin, Sergei. 1998. Extracting Patterns and Relations from the World Wide Web

2. Find Instances:
   - The Comedy of Error, by William Shakespeare, was…
   - The Comedy of Error, by William Shakespeare, is…
   - The Comedy of Error, one of William Shakespeare's earliest attempts…
   - The Comedy of Error, one of William Shakespeare's most…
3. Extract patterns (group by middle, take longest common prefix/suffix)
   - ?x, by ?y, was…
   - ?x, by ?y, is…
   - ?x, one of ?y's earliest attempts…
   - ?x, one of ?y's most…

APPROACHES

BOOTSTRAPPING

1. Start with 5 seeds:

Tuple: <author, book>

Author	Book
Isaac Asimov	The Robots of Dawn
David Grin	Startide Rising
James Gleick	Chaos: Making a New Science
Charles Dickems	Great Expectations
William Shakespeare	The Comedy of Errors

Brin, Sergei. 1998. Extracting Patterns and Relations from the World Wide Web

2. Find Instances:
   - The Comedy of Error, by William Shakespeare, was…
   - The Comedy of Error, by William Shakespeare, is…
   - The Comedy of Error, one of William Shakespeare's earliest attempts…
   - The Comedy of Error, one of William Shakespeare's most…
3. Extract patterns (group by middle, take longest common prefix/suffix)
   - ?x, by ?y, was…
   - ?x, by ?y, is…
   - ?x, one of ?y's earliest attempts…
   - ?x, one of ?y's most…
4. Iterate, finding new seeds that match the pattern

DISTANT SUPERVISION

INTUITION

… if two entities participate in a relation, any sentence that contains a pair of **entities** that participate in a known Freebase relation is likely to express that relation in some way.

↓

…extract very large numbers of features that are combined in a logistic regression classifier.

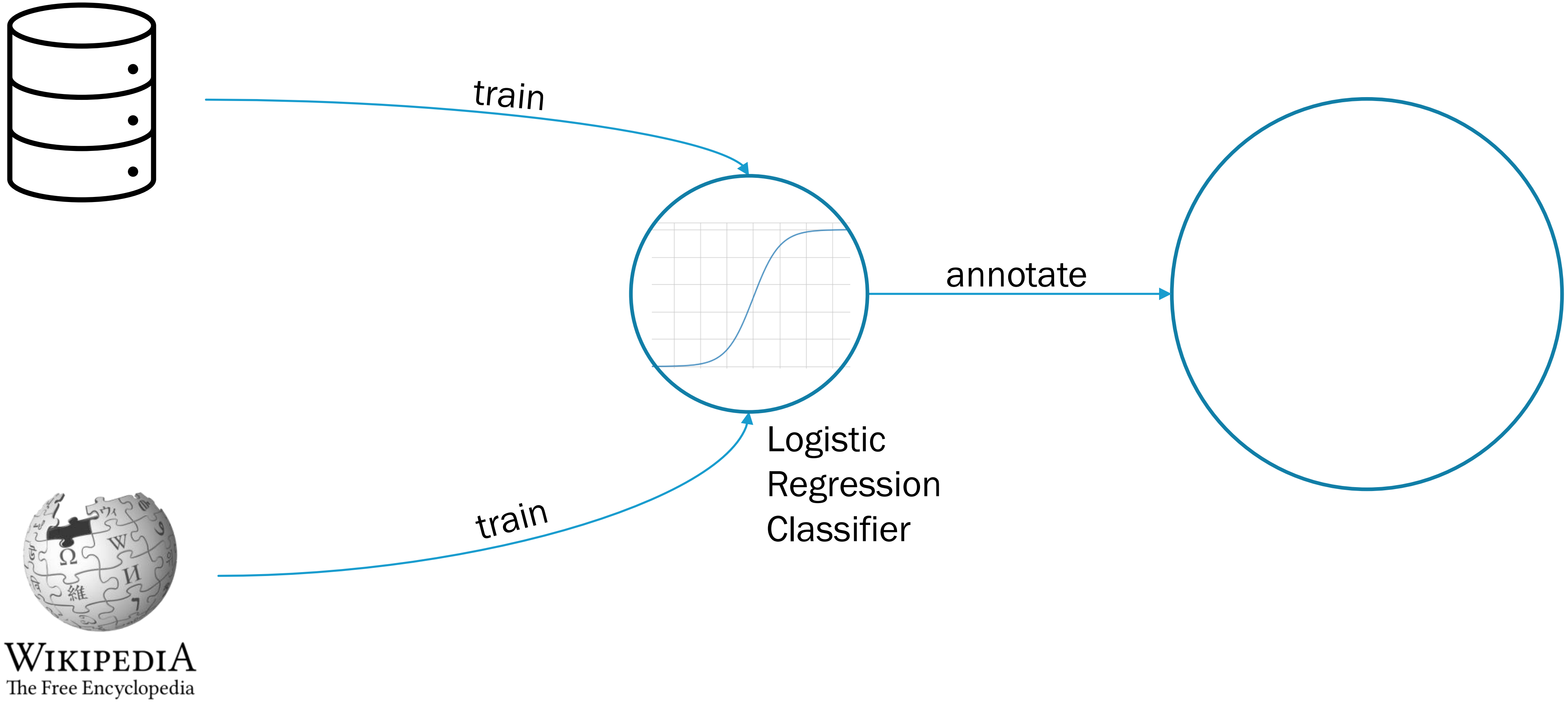
More text, more relations, and more instances.

DISTANT SUPERVISION

SUMMARY

Freebase

- 102 relations
- 940K entities
- 1.8 instances



DISTANT SUPERVISION
SUMMARY

- For each relation

born - in

DISTANT SUPERVISION

SUMMARY

- For each relation
- For each tuple in big database

study - in

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

DISTANT SUPERVISION
SUMMARY

- For each relation
- For each tuple in big database
- Find sentences in large corpus with both entities

born - in

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

Edwin Hubble was born in Marshfield
Einstein, born (1879), Ulm

DISTANT SUPERVISION
SUMMARY

- For each relation

born - in

- For each tuple in big database

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

- Find sentences in large corpus with both entities

Edwin Hubble was born in Marshfield
Einstein, born (1879), Ulm

- Extract frequent features

PER was born in LOC
PER, born (1879), LOC

DISTANT SUPERVISION

SUMMARY

- For each relation

born - in

- For each tuple in big database

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

- Find sentences in large corpus with both entities

Edwin Hubble was born in Marshfield
Einstein, born (1879), Ulm

- Extract frequent features

PER was born in LOC
PER, born (1879), LOC

- Train classifier

FEATURES

LEXICAL

	entity1	verb	verb	preposition	entity2		entity3
Astronomer	Edwin Hubble	was	born	in	Marshfield	,	Missouri

https://parts-of-speech.info/

Feature type	Left window	NE1	Middle	NE2	Right window
Lexical	[]	PER	[was/VERB born/VERB in/CLOSED]	LOC	[]
Lexical	[Astronomer]	PER	[was/VERB born/VERB in/CLOSED]	LOC	[,]
Lexical	[#PAD#, Astronomer]	PER	[was/VERB born/VERB in/CLOSED]	LOC	[, Missouri]
Syntactic	[]	PER	[$\Uparrow_s$ was $\Downarrow_{pred}$ born $\Downarrow_{mod}$ in $\Downarrow_{pcomp-n}$]	LOC	[]
Syntactic	[Edwin Hubble $\Downarrow_{lex-mod}$]	PER	[$\Uparrow_s$ was $\Downarrow_{pred}$ born $\Downarrow_{mod}$ in $\Downarrow_{pcomp-n}$]	LOC	[]
Syntactic	[Astronomer $\Downarrow_{lex-mod}$]	PER	[$\Uparrow_s$ was $\Downarrow_{pred}$ born $\Downarrow_{mod}$ in $\Downarrow_{pcomp-n}$]	LOC	[]
Syntactic	[]	PER	[$\Uparrow_s$ was $\Downarrow_{pred}$ born $\Downarrow_{mod}$ in $\Downarrow_{pcomp-n}$]	LOC	[$\Downarrow_{lex-mod}$,]
Syntactic	[Edwin Hubble $\Downarrow_{lex-mod}$]	PER	[$\Uparrow_s$ was $\Downarrow_{pred}$ born $\Downarrow_{mod}$ in $\Downarrow_{pcomp-n}$]	LOC	[$\Downarrow_{lex-mod}$,]
Syntactic	[Astronomer $\Downarrow_{lex-mod}$]	PER	[$\Uparrow_s$ was $\Downarrow_{pred}$ born $\Downarrow_{mod}$ in $\Downarrow_{pcomp-n}$]	LOC	[$\Downarrow_{lex-mod}$,]
Syntactic	[]	PER	[$\Uparrow_s$ was $\Downarrow_{pred}$ born $\Downarrow_{mod}$ in $\Downarrow_{pcomp-n}$]	LOC	[$\Downarrow_{inside}$ Missouri]
Syntactic	[Edwin Hubble $\Downarrow_{lex-mod}$]	PER	[$\Uparrow_s$ was $\Downarrow_{pred}$ born $\Downarrow_{mod}$ in $\Downarrow_{pcomp-n}$]	LOC	[$\Downarrow_{inside}$ Missouri]
Syntactic	[Astronomer $\Downarrow_{lex-mod}$]	PER	[$\Uparrow_s$ was $\Downarrow_{pred}$ born $\Downarrow_{mod}$ in $\Downarrow_{pcomp-n}$]	LOC	[$\Downarrow_{inside}$ Missouri]

Variations*:

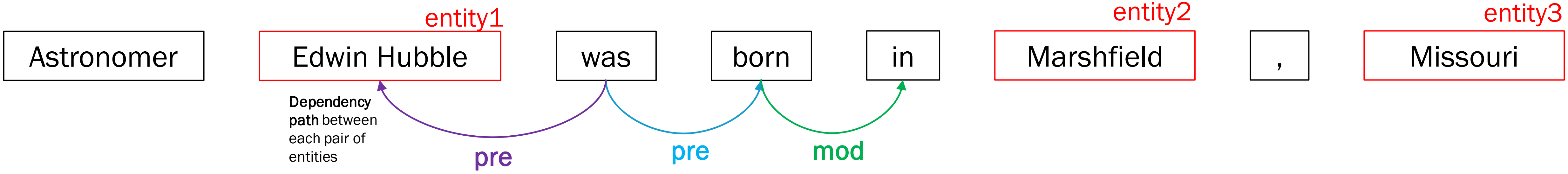
- omitting all words that are not verbs,
- omitting all function words.

Precision increment, but not large enough to justify the computational effort.

*attempt to approximate syntactic features

FEATURES

SYNTACTIC



Feature type	Left window	NE1	Middle	NE2	Right window
Lexical	[]	PER	[was/VERB born/VERB in/CLOSED]	LOC	[]
Lexical	[Astronomer]	PER	[was/VERB born/VERB in/CLOSED]	LOC	[,]
Lexical	[#PAD#, Astronomer]	PER	[was/VERB born/VERB in/CLOSED]	LOC	[, Missouri]
Syntactic	[]	PER	[$\Uparrow_s$ was $\Downarrow_{pred}$ born $\Downarrow_{mod}$ in $\Downarrow_{pcomp-n}$]	LOC	[]
Syntactic	[Edwin Hubble $\Downarrow_{lex-mod}$]	PER	[$\Uparrow_s$ was $\Downarrow_{pred}$ born $\Downarrow_{mod}$ in $\Downarrow_{pcomp-n}$]	LOC	[]
Syntactic	[Astronomer $\Downarrow_{lex-mod}$]	PER	[$\Uparrow_s$ was $\Downarrow_{pred}$ born $\Downarrow_{mod}$ in $\Downarrow_{pcomp-n}$]	LOC	[]
Syntactic	[]	PER	[$\Uparrow_s$ was $\Downarrow_{pred}$ born $\Downarrow_{mod}$ in $\Downarrow_{pcomp-n}$]	LOC	[$\Downarrow_{lex-mod}$,]
Syntactic	[Edwin Hubble $\Downarrow_{lex-mod}$]	PER	[$\Uparrow_s$ was $\Downarrow_{pred}$ born $\Downarrow_{mod}$ in $\Downarrow_{pcomp-n}$]	LOC	[$\Downarrow_{lex-mod}$,]
Syntactic	[Astronomer $\Downarrow_{lex-mod}$]	PER	[$\Uparrow_s$ was $\Downarrow_{pred}$ born $\Downarrow_{mod}$ in $\Downarrow_{pcomp-n}$]	LOC	[$\Downarrow_{lex-mod}$,]
Syntactic	[]	PER	[$\Uparrow_s$ was $\Downarrow_{pred}$ born $\Downarrow_{mod}$ in $\Downarrow_{pcomp-n}$]	LOC	[$\Downarrow_{inside}$ Missouri]
Syntactic	[Edwin Hubble $\Downarrow_{lex-mod}$]	PER	[$\Uparrow_s$ was $\Downarrow_{pred}$ born $\Downarrow_{mod}$ in $\Downarrow_{pcomp-n}$]	LOC	[$\Downarrow_{inside}$ Missouri]
Syntactic	[Astronomer $\Downarrow_{lex-mod}$]	PER	[$\Uparrow_s$ was $\Downarrow_{pred}$ born $\Downarrow_{mod}$ in $\Downarrow_{pcomp-n}$]	LOC	[$\Downarrow_{inside}$ Missouri]

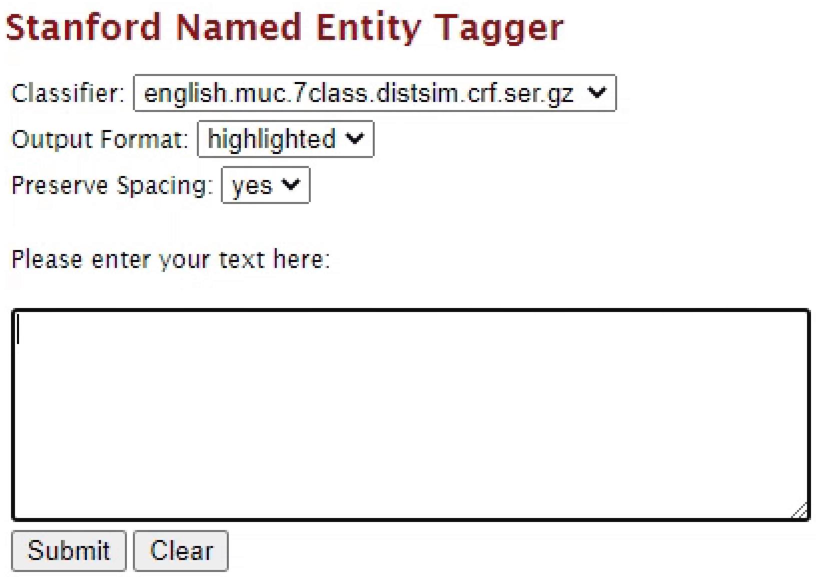
Grammatical Relationships

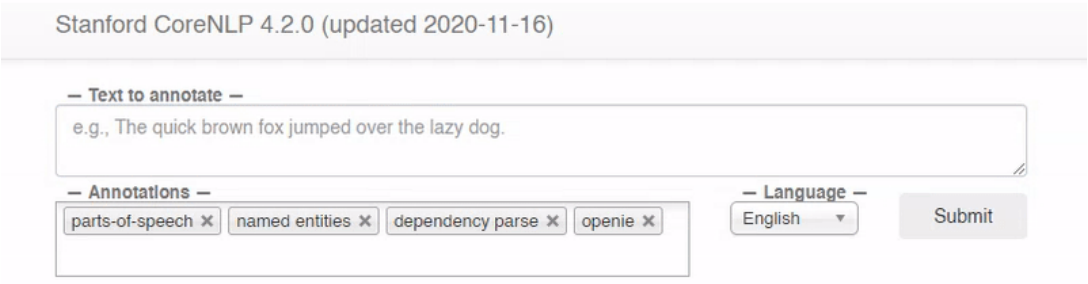
Symbol	Description
inv-aux	inverted auxiliary
mod	relationship between a word and its adjunct modifier
pnmod	post nominal modifier
p-spec	specifier of prepositional phrases
pcomp-c	clausal complement of prepositions
pcomp-n	nominal complement of prepositions
post	post determiner
pre	pre determiner
s	surface subjec

https://gate.ac.uk/releases/gate-7.0-build4195-ALL/doc/tao/splitch17.html

FEATURES

NAMED ENTITY TAGS - STANFORD NAMED ENTITY TAGGER





Reference:

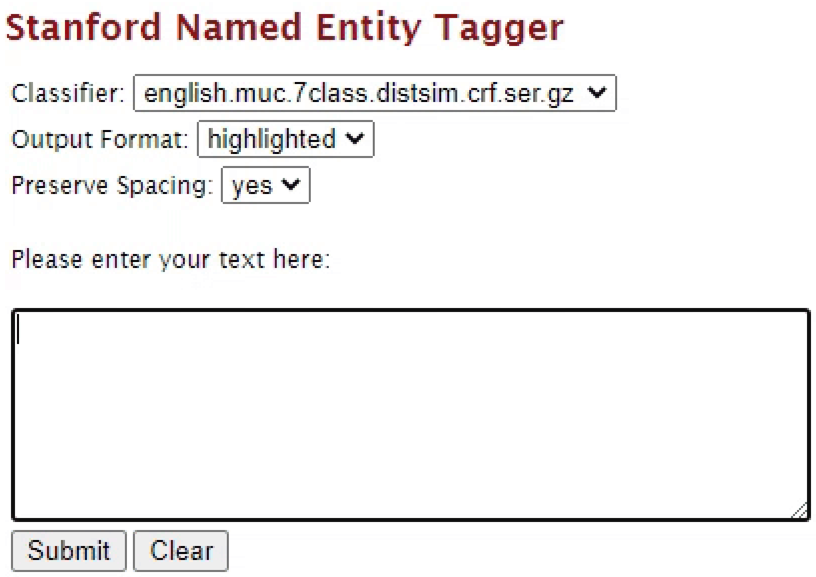
- http://nlp.stanford.edu:8080/ner/

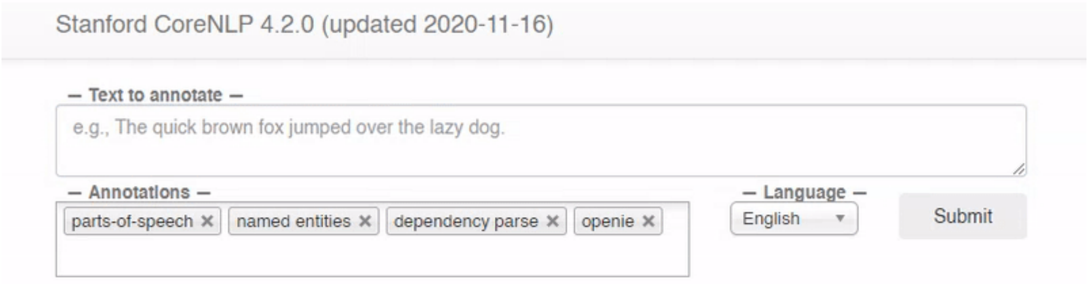
Reference:

- http://corenlp.run/

FEATURES

NAMED ENTITY TAGS - STANFORD NAMED ENTITY TAGGER





Reference:

- http://nlp.stanford.edu:8080/ner/

Reference:

- http://corenlp.run/

FEATURES

FEATURE CONJUNCTIONS

Example

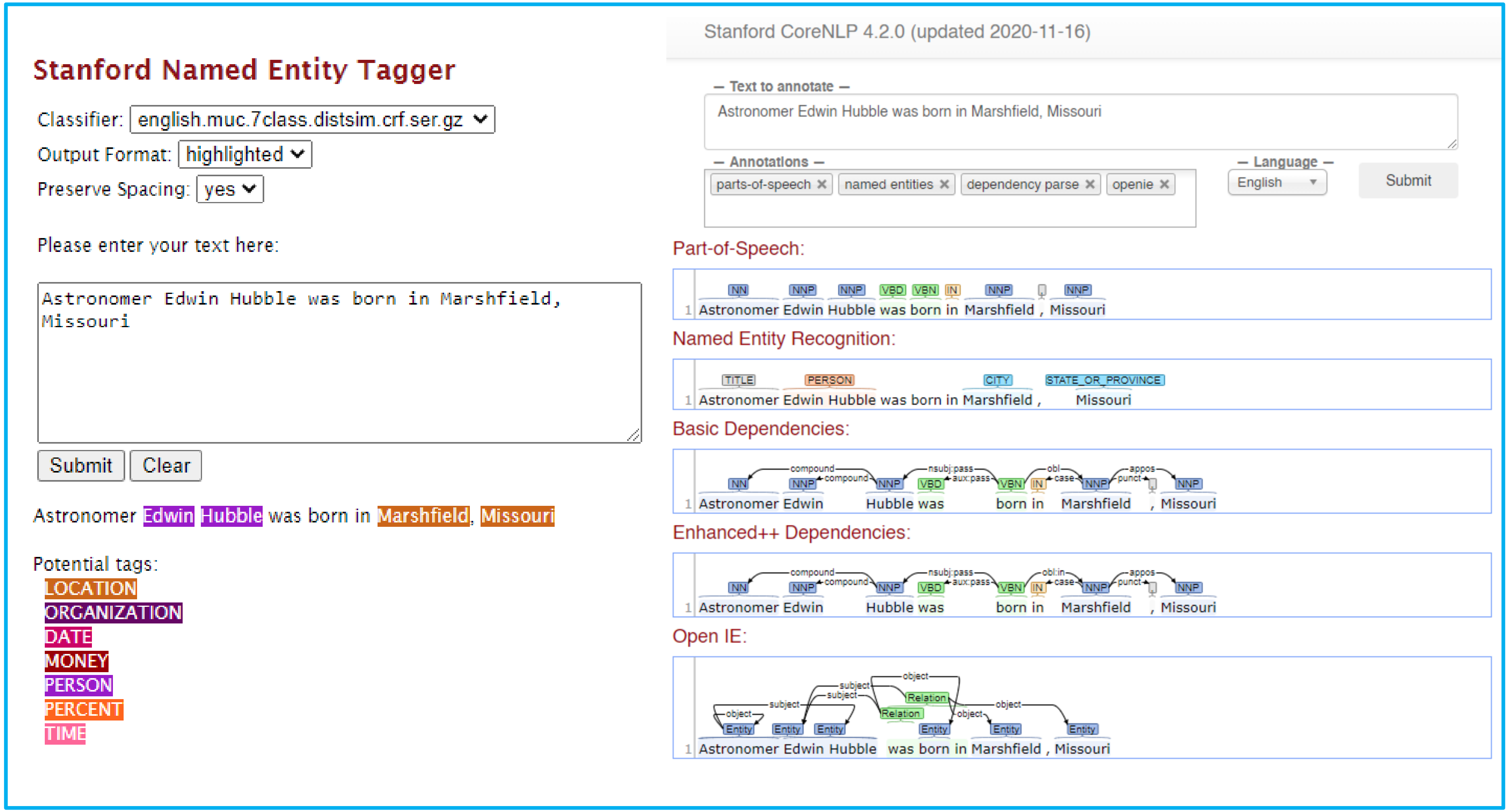
Reference:

- http://corenlp.run/

FEATURES

FEATURE CONJUNCTIONS

Named Entity Tags



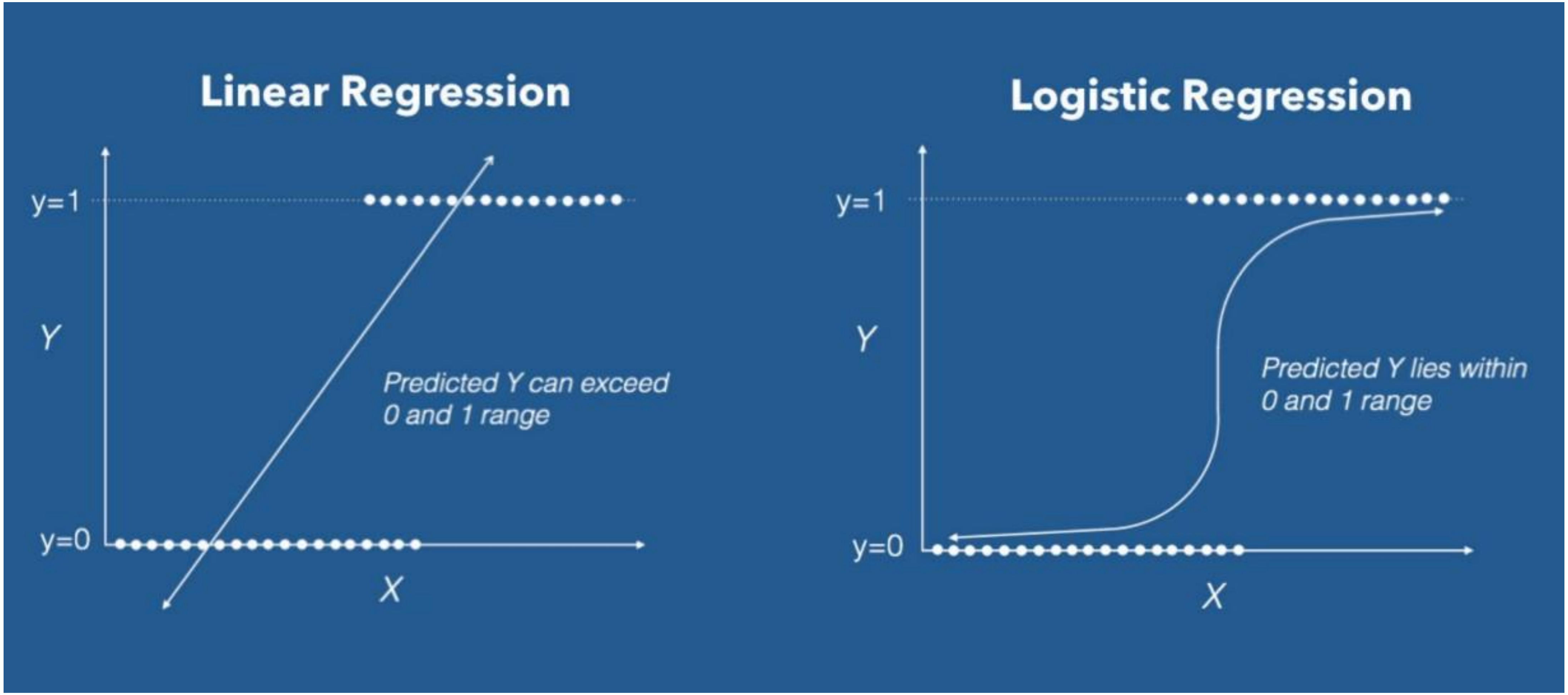
+ Attributes in sentence

↓ Recall ↑ Precision

Large dataset

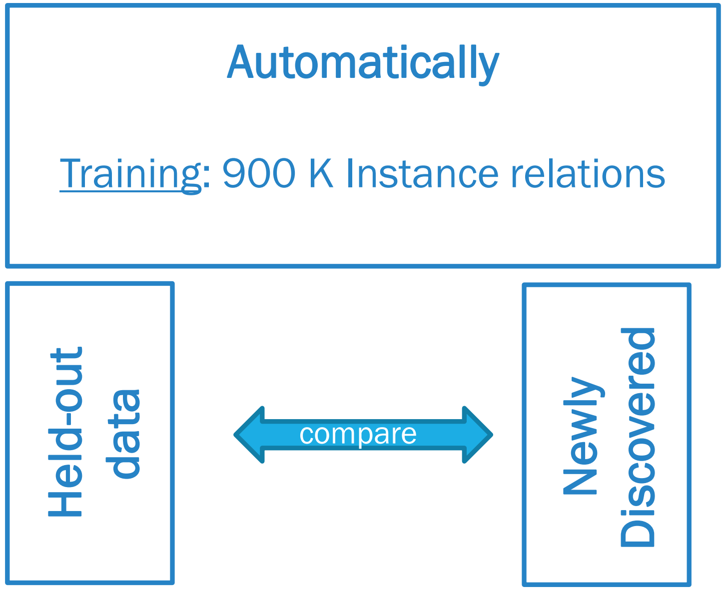
IMPLEMENTATION

MULTI-CLASS LOGISTIC CLASSIFIER OPTIMIZED USING L-BFGS WITH GAUSSIAN REGULARIZATION



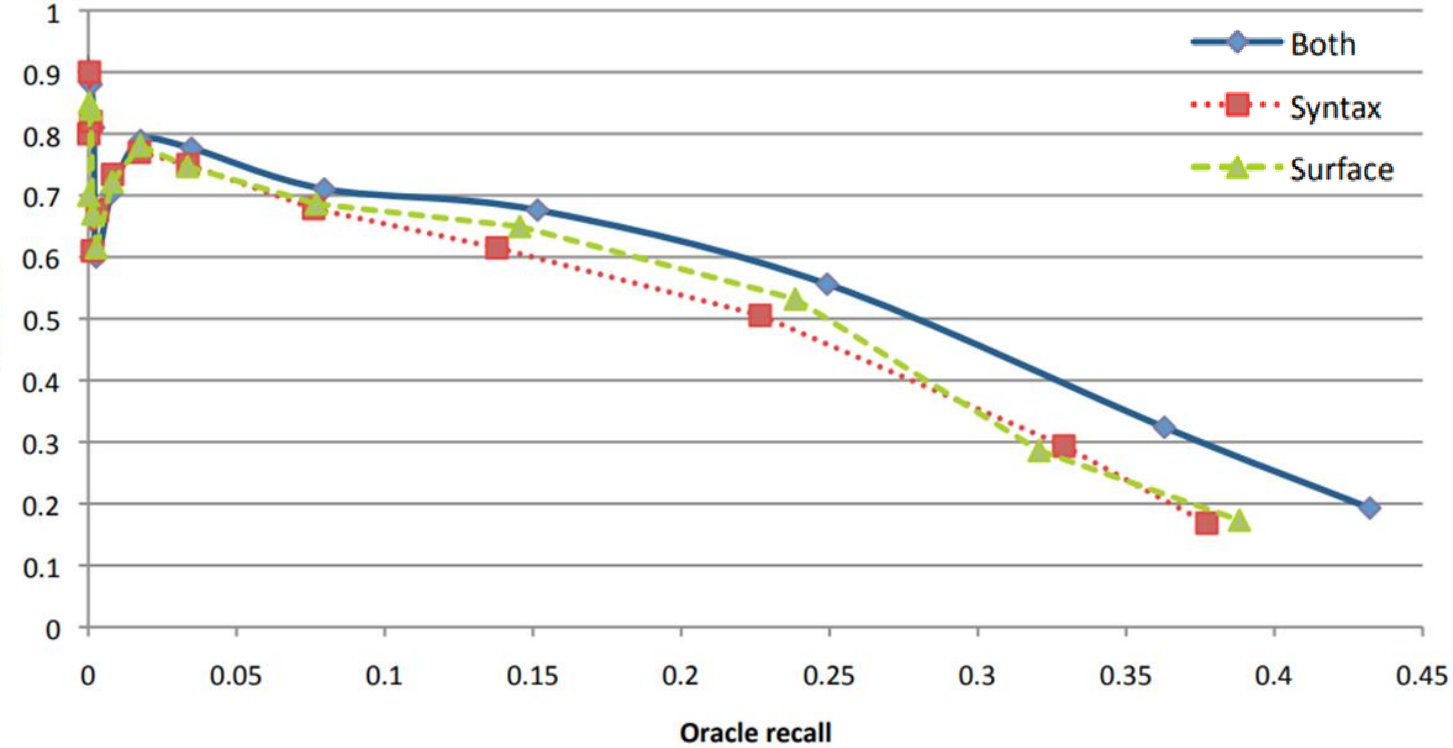
EVALUATION

AUTOMATIC – HELD-OUT



		Real Label	
		Positive	Negative
Predicted Label	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

Combination of syntactic and lexical features provides better performance than either feature set on its own.



$$\text{Precision} = \frac{\sum \text{TP}}{\sum \text{TP} + \text{FP}}$$

$$\text{Recall} = \frac{\sum \text{TP}}{\sum \text{TP} + \text{FN}}$$

- The combination of syntactic and lexical features offers a substantial improvement in precision over either of these feature sets on its own.
- At the 100,000-recall level, we classify most of the instances into three relations:
  - 60% as location-contains
  - 13% as person-place-of-birth
  - 10% as person-nationality

Automatic evaluation with 50% of Freebase relation data held out and 50% used in training on the 102 largest relations we use.

[Source: Mintz et al., 2009, Distant supervision for relation extraction without labeled data

EVALUATION

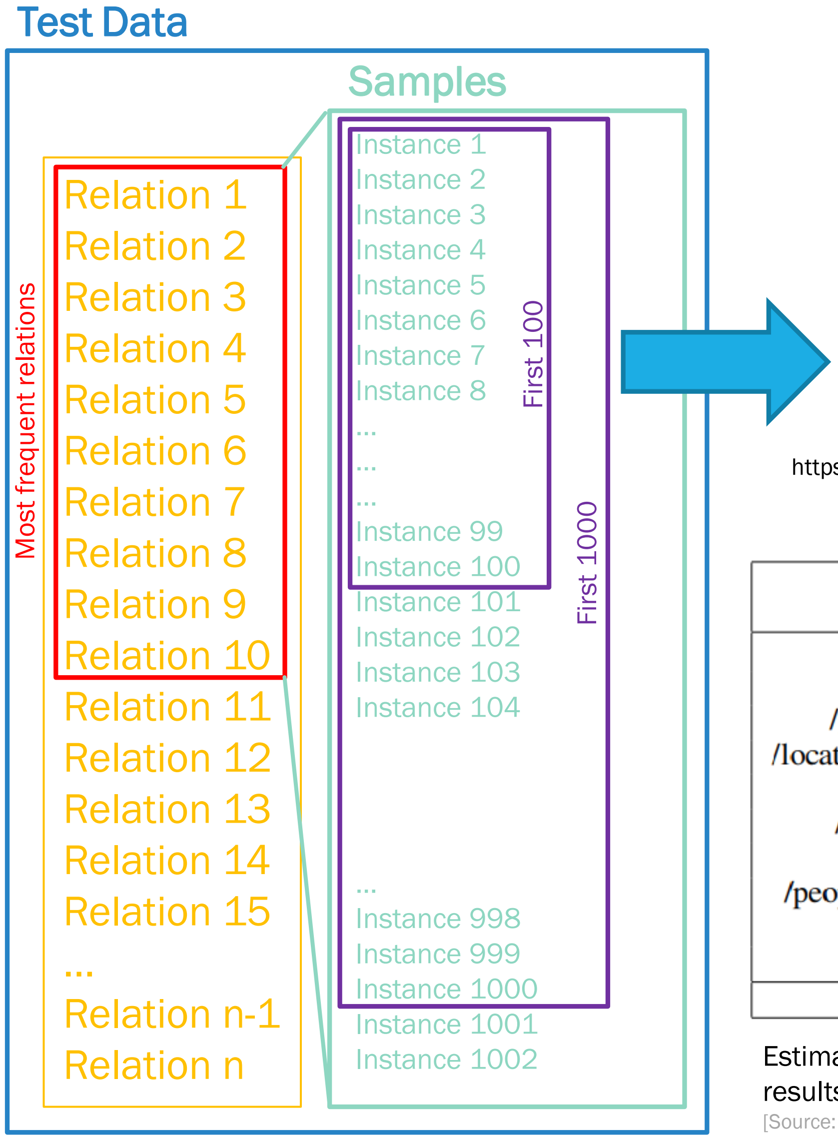
HUMAN

Manually

Syntactic features

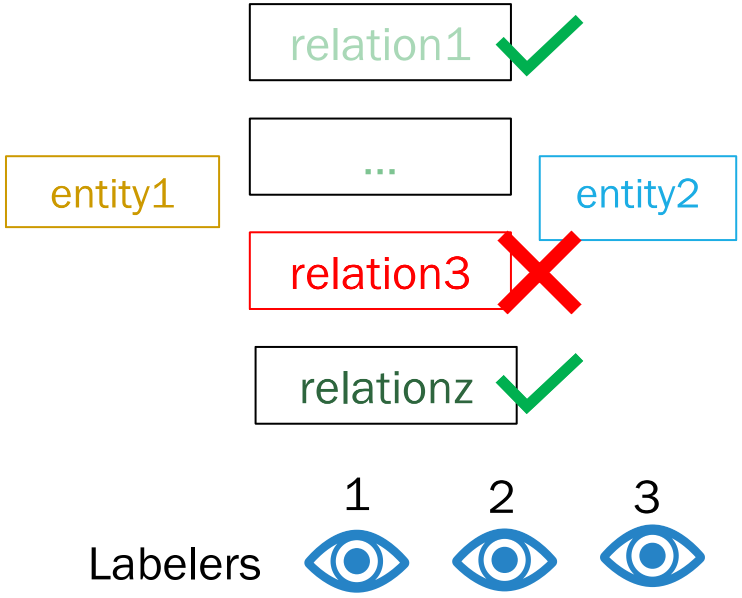
Lexical features

Syntactic + Lexical features



Amazon's Mechanical Turk service

https://www.mturk.com/



Relation name	100 instances Syn	100 instances Lex	100 instances Both	1000 instances Syn	1000 instances Lex	1000 instances Both
/film/director/film	**0.49**	0.43	0.44	**0.49**	0.41	0.46
/film/writer/film	**0.70**	0.60	0.65	**0.71**	0.61	0.69
/geography/river/basin_countries	0.65	0.64	**0.67**	**0.73**	0.71	0.64
/location/country/administrative_divisions	0.68	0.59	**0.70**	**0.72**	0.68	**0.72**
/location/location/contains	0.81	**0.89**	0.84	**0.85**	0.83	0.84
/location/us_county/county_seat	0.51	0.51	**0.53**	0.47	**0.57**	0.42
/music/artist/origin	0.64	0.66	**0.71**	0.61	**0.63**	0.60
/people/deceased_person/place_of_death	0.80	0.79	**0.81**	0.80	**0.81**	0.78
/people/person/nationality	0.61	0.70	**0.72**	0.56	0.61	**0.63**
/people/person/place_of_birth	**0.78**	0.77	**0.78**	0.88	0.85	**0.91**
Average	0.67	0.66	**0.69**	**0.68**	0.67	0.67

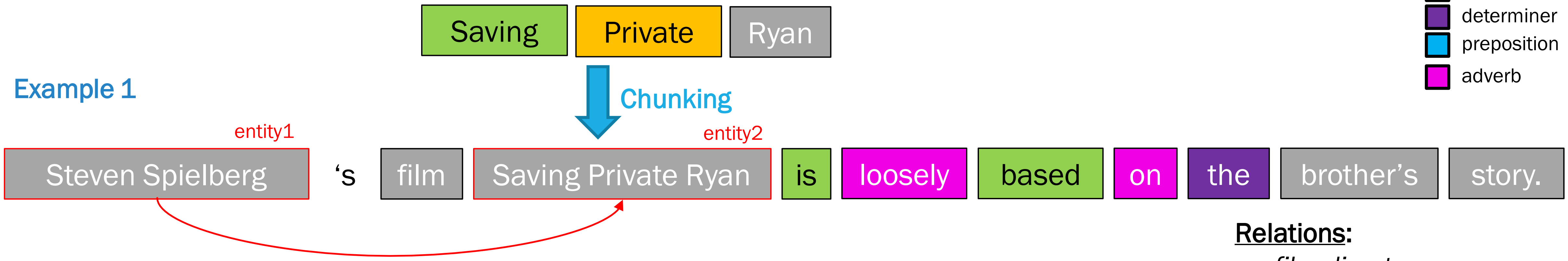
Estimated precision on human-evaluation experiments of the highest-ranked 100 and 1000 results per relation, using stratified samples..

[Source: Mintz et al., 2009, Distant supervision for relation extraction without labeled data

RESULTS DISCUSSION

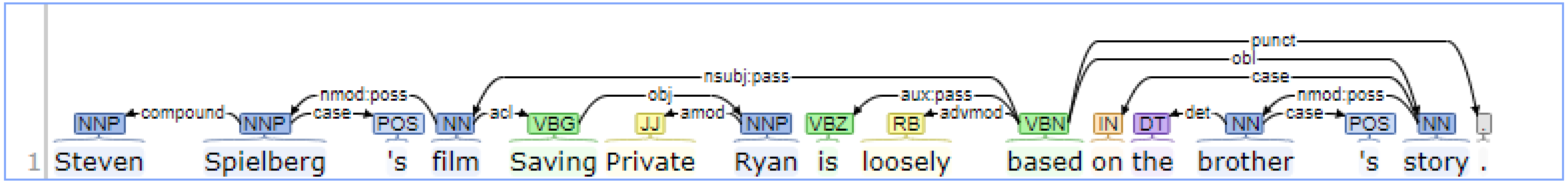
noun
verb
determiner
preposition
adverb

Example 1



Relations:

- *film-director*
- *film-writer*
- *film-producer*



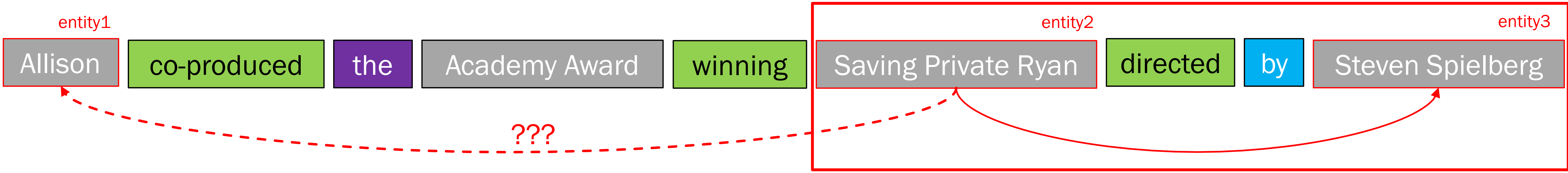
RESULTS DISCUSSION

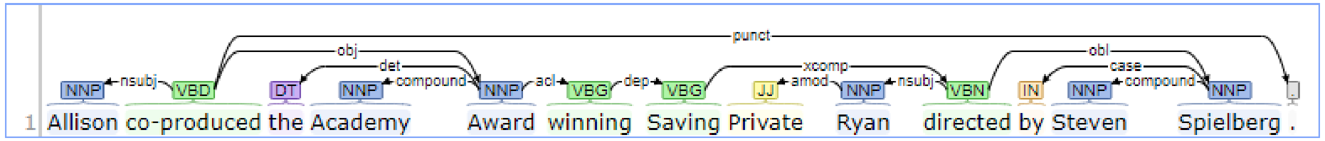
Example 2

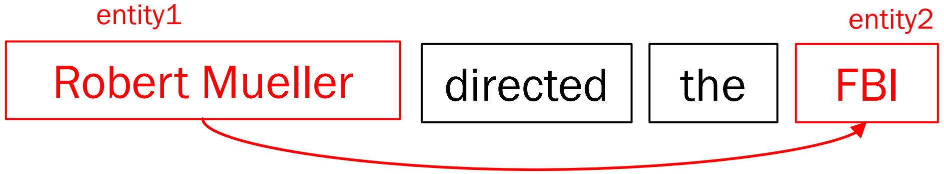
What is ???

- Film?
- TV show?
- Book?

noun
verb
determiner
preposition
adverb







Relations:

- *film-director*
- *CEO-company*

Relation name	100 instances			1000 instances		
	Syn	Lex	Both	Syn	Lex	Both
/film/director/film	**0.49**	0.43	0.44	**0.49**	0.41	0.46
/film/writer/film	**0.70**	0.60	0.65	**0.71**	0.61	0.69
/geography/river/basin_countries	0.65	0.64	**0.67**	**0.73**	0.71	0.64
/location/country/administrative_divisions	0.68	0.59	**0.70**	**0.72**	0.68	**0.72**
/location/location/contains	0.81	**0.89**	0.84	**0.85**	0.83	0.84
/location/us_county/county_seat	0.51	0.51	**0.53**	0.47	**0.57**	0.42
/music/artist/origin	0.64	0.66	**0.71**	0.61	**0.63**	0.60
/people/deceased_person/place_of_death	0.80	0.79	**0.81**	0.80	**0.81**	0.78
/people/person/nationality	0.61	0.70	**0.72**	0.56	0.61	**0.63**
/people/person/place_of_birth	**0.78**	0.77	**0.78**	0.88	0.85	**0.91**
Average	0.67	0.66	**0.69**	**0.68**	0.67	0.67

Estimated precision on human-evaluation experiments of the highest-ranked 100 and 1000 results per relation, using stratified samples..

[Source: Mintz et al., 2009, Distant supervision for relation extraction without labeled data

In case of ambiguous relations,

syntactic features are **important.**

CONCLUSIONS

Distant supervision for relation extraction without labeled data (2009)

Authors: Mike Mintz, Steven Bills, Rion Snow, Dan Jurafsky

- Extract relations from unlabeled text
- Using database, the label is suit for the current database
- Extracted relations has a 67% of accuracy.

DISCUSSION

Distant supervision for relation extraction without labeled data (2009)

Authors: Mike Mintz, Steven Bills, Rion Snow, Dan Jurafsky

- **Distantly supervision** can be useful for other taks.
- It is necessary to have a large database.

THANKS

QUESTIONS?

APPROACHES
MANUAL

PER GPE

Roderick Perez was in Colombia → Person - Located

SUMMARY AND KEY POINTS

Distant supervision for relation extraction without labeled data (2009)

Authors: Mike Mintz, Steven Bills, Rion Snow, Dan Jurafsky

- New sentence (using NER):

entity1					entity2
Vienna	is	the	capital	of	Austria

- Check for sentences which **Named Entity Tagged** exist in a known Freebase relation.

FreeBase

Bogota is the capital of **Colombia**

Richmond, is the capital of **Virginia**

Caracas is the capital of **Venezuela**

...

Henry's of Nantes helped ... in **France**

location – contains
film – director
film – writer
music – art - origin
...
people/profession

- Since there is a large dataset, many sentence might cover such relation (larger text, more potential features).

- Use a Logistic classifier to map the relation

If two entities have a known relation in a knowledge base, all sentences that mention these two entities will probably express the same relation.

DISTANT SUPERVISION

SUMMARY

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

IMPLEMENTATION TEXT

Freebase Triples

This dataset contains every fact currently in Freebase.

Total triples: 1.9 billion
Updated: Weekly
Data Format: N-Triples RDF
License: CC-BY

22 GB gzip
250 GB uncompressed

Download

https://developers.google.com/freebase/

Unstructured

Wikipedia*

Articles: 1.8 M
Sentence per article: 14.3 / article
Words: 601,600,703

Training

800 K

Testing

400 K

*not including discussion or user pages.

Freebase Wikipedia Extraction

https://github.com/saleiro/Freebase-to-Wikipedia

Training: 900 K Instance relations**

** originally 1.8 M (only used for human evaluation experiments during training).

Structured

Relation 1
Relation 2
X study at Y
…
Relation n-1
Relation n

NEW relation

Feature 1
Feature 2
Feature 3
…
Feature n-1
Feature n

Entity 1
Entity 2
Entity 3
…
Entity n-1
Entity n

extract relation instances that do not appear in our training data (not already in Freebase).

FREEBASE-WIKIPEDIA

Wikipedia

Wikipedia Infobox

Relations

ORG GPE

UniWien location Vienna

Organization - Located

CLASSIFICATION RESULTS

Relation	Feature type	Left window	NE1	Middle	NE2	Right window
/architecture/structure/architect	LEX↶		ORG	, the designer of the	PER	
	SYN	designed $\Uparrow_{s}$	ORG	$\Uparrow_{s}$ designed $\Downarrow_{by-subj}$ by $\Downarrow_{pcn}$	PER	$\Uparrow_{s}$ designed
/book/author/works_written	LEX		PER	s novel	ORG	
	SYN		PER	$\Uparrow_{pcn}$ by $\Uparrow_{mod}$ story $\Uparrow_{pred}$ is $\Downarrow_{s}$	ORG	
/book/book_edition/author_editor	LEX↶		ORG	s novel	PER	
	SYN		PER	$\Uparrow_{nn}$ series $\Downarrow_{gen}$	PER	
/business/company/founders	LEX		ORG	co - founder	PER	
	SYN		ORG	$\Uparrow_{nn}$ owner $\Downarrow_{person}$	PER	
/business/company/place_founded	LEX↶		ORG	- based	LOC	
	SYN		ORG	$\Uparrow_{s}$ founded $\Downarrow_{mod}$ in $\Downarrow_{pcn}$	LOC	
/film/film/country	LEX		PER	, released in	LOC	
	SYN	opened $\Uparrow_{s}$	ORG	$\Uparrow_{s}$ opened $\Downarrow_{mod}$ in $\Downarrow_{pcn}$	LOC	$\Uparrow_{s}$ opened
/geography/river/mouth	LEX		LOC	, which flows into the	LOC	
	SYN	the $\Downarrow_{det}$	LOC	$\Uparrow_{s}$ is $\Downarrow_{pred}$ tributary $\Downarrow_{mod}$ of $\Downarrow_{pcn}$	LOC	$\Downarrow_{det}$ the
/government/political_party/country	LEX↶		ORG	politician of the	LOC	
	SYN	candidate $\Uparrow_{nn}$	ORG	$\Uparrow_{nn}$ candidate $\Downarrow_{mod}$ for $\Downarrow_{pcn}$	LOC	$\Uparrow_{nn}$ candidate
/influence/influence_node/influenced	LEX↶		PER	, a student of	PER	
	SYN	of $\Uparrow_{pcn}$	PER	$\Uparrow_{pcn}$ of $\Uparrow_{mod}$ student $\Uparrow_{appo}$	PER	$\Uparrow_{pcn}$ of
/language/human_language/region	LEX		LOC	- speaking areas of	LOC	
	SYN		LOC	$\Uparrow_{lex-mod}$ speaking areas $\Downarrow_{mod}$ of $\Downarrow_{pcn}$	LOC	
/music/artist/origin	LEX↶		ORG	based band	LOC	
	SYN	is $\Uparrow_{s}$	ORG	$\Uparrow_{s}$ is $\Downarrow_{pred}$ band $\Downarrow_{mod}$ from $\Downarrow_{pcn}$	LOC	$\Uparrow_{s}$ is
/people/deceased_person/place_of_death	LEX		PER	died in	LOC	
	SYN	hanged $\Uparrow_{s}$	PER	$\Uparrow_{s}$ hanged $\Downarrow_{mod}$ in $\Downarrow_{pcn}$	LOC	$\Uparrow_{s}$ hanged
/people/person/nationality	LEX		PER	is a citizen of	LOC	
	SYN		PER	$\Downarrow_{mod}$ from $\Downarrow_{pcn}$	LOC	
/people/person/parents	LEX		PER	, son of	PER	
	SYN	father $\Uparrow_{gen}$	PER	$\Uparrow_{gen}$ father $\Downarrow_{person}$	PER	$\Uparrow_{gen}$ father
/people/person/place_of_birth	LEX↶		PER	is the birthplace of	PER	
	SYN		PER	$\Uparrow_{s}$ born $\Downarrow_{mod}$ in $\Downarrow_{pcn}$	LOC	
/people/person/religion	LEX		PER	embraced	LOC	
	SYN	convert $\Downarrow_{appo}$	PER	$\Downarrow_{appo}$ convert $\Downarrow_{mod}$ to $\Downarrow_{pcn}$	LOC	$\Downarrow_{appo}$ convert

SYN syntactic feature

LEX lexical feature

reversed

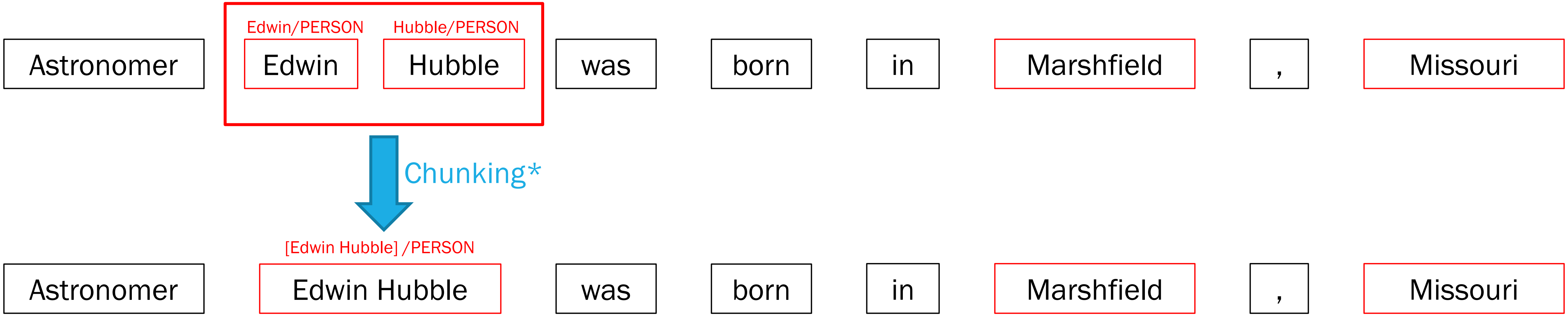
NE# named entity tag of entity

Examples of high-weight features for several relations

[Source: Mintz et al., 2009, Distant supervision for relation extraction without labeled data

IMPLEMENTATION
PARSING AND CHUNKING

PreProcessing

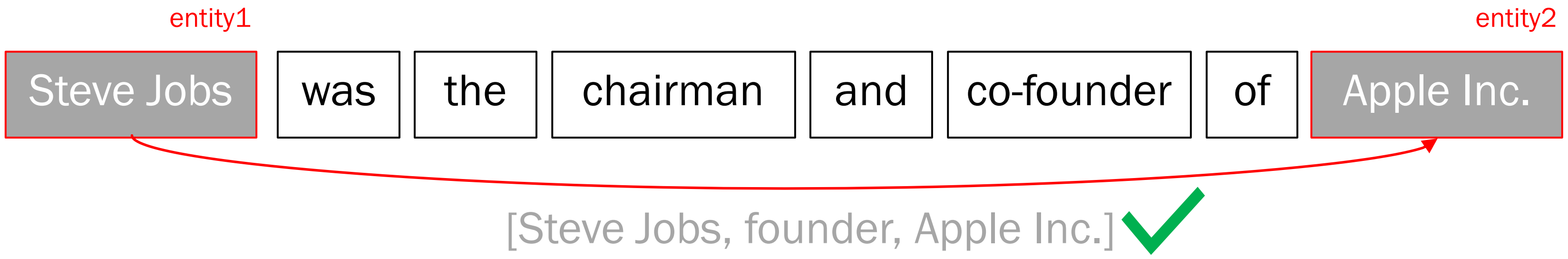


*restricted by the dependency parse of the sentence.

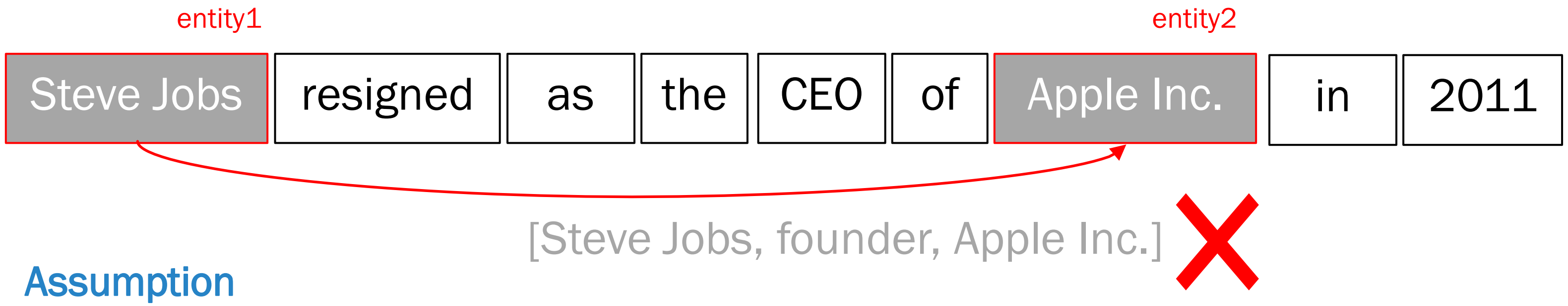
RECENT WORK

GAN Driven Semi-distant Supervision for Relation Extraction

<u>Author</u>: Shanghai Jiao Tong University



False Positive Intances



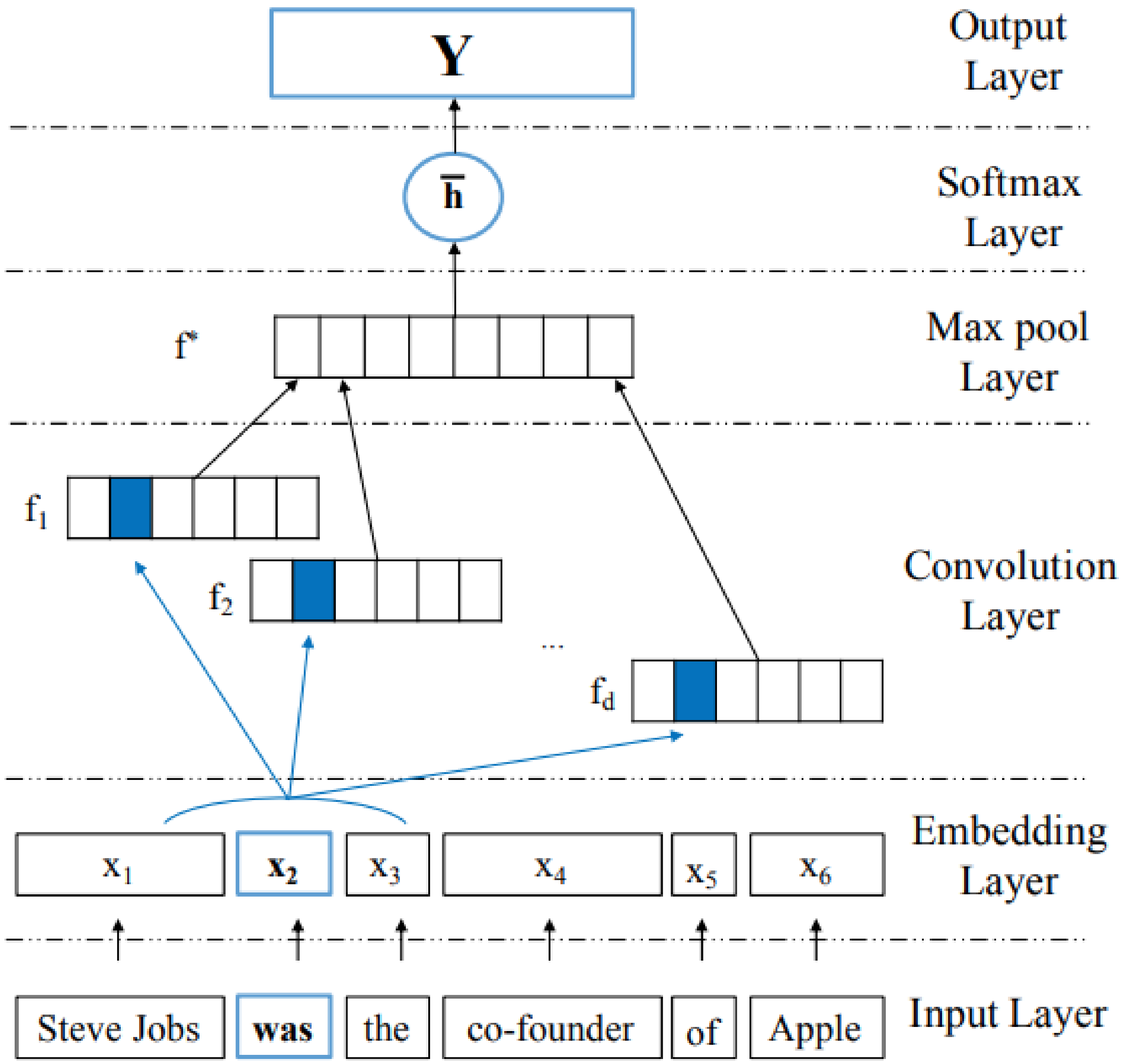
Assumption

If an entity relevant to another entity, its name is possibly mentioned in the description of the other entity.

RECENT WORK

GAN Driven Semi-distant Supervision for Relation Extraction

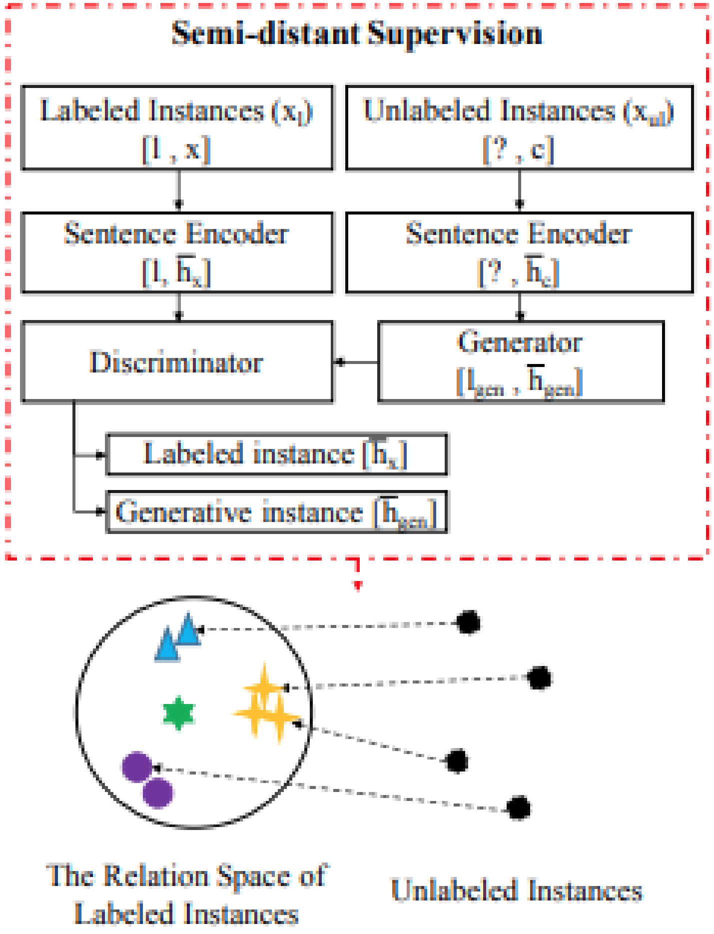
Author: Shanghai Jiao Tong University



RECENT WORK

GAN Driven Semi-distant Supervision for Relation Extraction

Author: Shanghai Jiao Tong University



Datasets

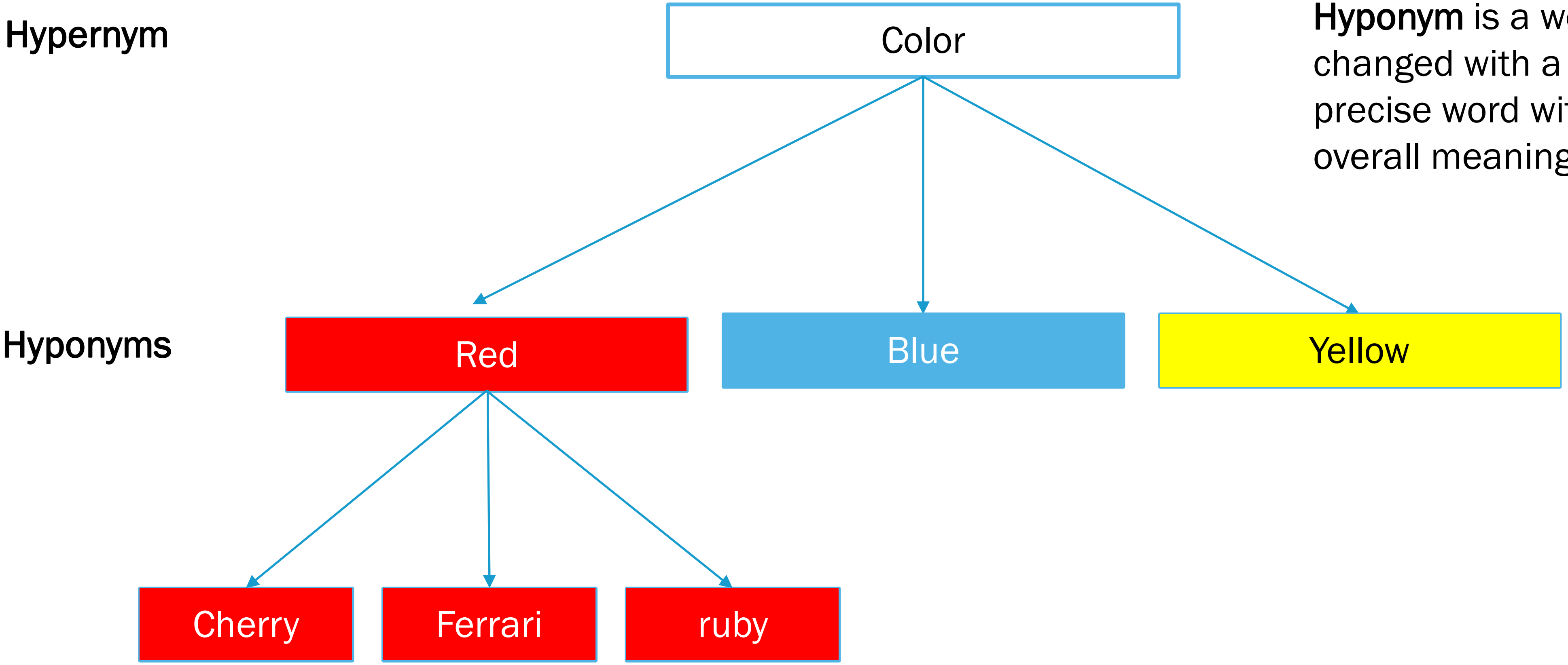
Dataset	Positive	Negative	Unlabeled	Classes
NYT	163,108	579,428	-	53
A-NYT	163,108	240,453	338,975	53
Books	1,000	1,000	2,000	2
Electronics	1,000	998	2,000	2
DVD	1,000	1,000	2,000	2
Kitchen	1,000	1,000	2,000	2
Apparel	1,000	1,000	2,000	2
Camera	999	998	2,000	2
Health	1,000	1,000	2,000	2
Music	1,000	1,000	2,000	2
Toys	1,000	1,000	2,000	2
Video	1,000	1,000	2,000	2
Baby	1,000	900	2,000	2
Magazine	1,000	970	2,000	2
Software	1,000	915	475	2
Sports	1,000	1,000	2,000	2
IMDB	994	1,006	2,000	2
MR	986	1,014	2,000	2

Results

P@N	100	200	300	Mean	PR
Zeng et al. (2015)	72.3	69.7	64.1	68.7	0.33
Lin et al. (2016)	76.2	73.1	67.4	72.2	0.35
Wu et al. (2017)	81.0	74.5	71.7	75.7	0.34
Liu et al. (2017b)	87.0	84.5	77.0	82.8	0.34
Qin et al. (2018a)	78.0	75.5	72.3	75.3	0.35
Liu et al. (2018)	87.0	83.0	78.0	82.7	0.39
Our Method	**96.0**	**93.5**	**93.0**	**94.2**	**0.56**

DISTANT SUPERVISION

HYPERNYM (IS – A)



Hyponym is a word that can be changed with a different and less precise word without changing the overall meaning of the phrase.[1]

Source: https://simple.wikipedia.org/wiki/Hyponymy_and_hypernymy

RELATION EXTRACTION

Relation name	New instance
/location/location/contains	Paris, Montmartre
/location/location/contains	Ontario, Fort Erie
/music/artist/origin	Mighty Wagon, Cincinnati
/people/deceased_person/place_of_death	Fyodor Kamensky, Clearwater
/people/person/nationality	Marianne Yvonne Heemskerk, Netherlands
/people/person/place_of_birth	Wavell Wayne Hinds, Kingston
/book/author/works_written	Upton Sinclair, Lanny Budd
/business/company/founders	WWE, Vince McMahon
/people/person/profession	Thomas Mellon, judge

Ten relation instances extracted by the system that did not appear in Freebase

[Source: Mintz et al., 2009, Distant supervision for relation extraction without labeled data