Information extraction & automated knowledge graph construction

Lecture 2: Entities & relations

Lauriane Aufrant (Inria Paris)

Université Paris Cité – UFRL – M2 LI – 16/11/2023

Progress

- ~~09/11 – Overview of information extraction~~
- 16/11 – Entities, relations...
- 23/11 – Coreference, linking...
- 30/11 – From IE to automated knowledge graph construction
- 07/12 – IE annotations
- 14/12 – IE in a specialty domain

Today: Entities & relations

- Named entity recognition

  $\hookrightarrow$ task, formats, corpora, methods, tools, evaluation

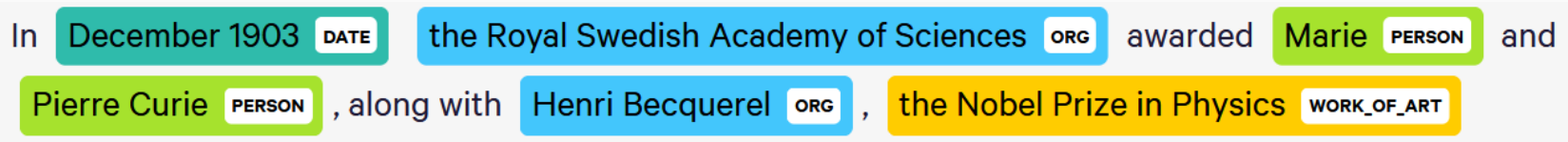
- Relation extraction

- N-ary relation extraction

Named entity recognition

"In December 1903 the Royal Swedish Academy of Sciences awarded Marie and Pierre Curie, along with Henri Becquerel, the Nobel Prize in Physics"

From spaCy:



Named Entity Recognition (NER)

Detecting and categorizing mentions of names according to the entity's type

MUC-7 Named Entity Task Definition

Version 3.5

17 September 1997

Nancy Chinchor (chinchor@gso.saic.com)

1. INTRODUCTION

1.1 Scope

The Named Entity task consists of three subtasks (entity names, temporal expressions, number expressions). The expressions to be annotated are "unique identifiers" of entities (organizations, persons, locations), times (dates, times), and quantities (monetary values, percentages).

For many text processing systems, such identifiers are recognized primarily using local pattern-matching techniques. The TEI (Text Encoding Initiative) Guidelines for Electronic Text Encoding and Interchange cover such identifiers (plus abbreviations) together in section 6.4 and explain that the identifiers comprise "textual features which it is often convenient to distinguish from their surrounding text. Names, dates and numbers are likely to be of particular importance to the scholar treating a text as source for a database; distinguishing such items from the surrounding text is however equally important to the scholar primarily interested in lexis."

The task is to identify all instances of the three types of expressions in each text in the test set and to subcategorize the expressions. The original texts contain some SGML tags already; the Named Entity task is to be performed within the text delimited by the SLUG, DATE, NWORDS, PREAMBLE, TEXT, and TRAILER tags.

The system must produce a single, unambiguous output for any relevant string in the text; thus, this evaluation is not based on a view of a pipelined system architecture in which Named Entity recognition would be completely handled as a preprocess to sentence and discourse analysis. The task requires that the system recognize what a string represents, not just its superficial appearance. Sometimes, the right answer is superficially apparent, as in the case of most, if not all, NUMEX expressions, and can be obtained by local pattern-matching techniques. In other cases, the right answer is not superficially apparent, as when a single capitalized word could represent the name of a location, person, or organization, and the answer may have to be obtained using techniques that draw information from a larger context or from reference lists.

The three subtasks correspond to three SGML tag elements: ENAMEX, TIMEX, and NUMEX. The subcategorization is captured by a SGML tag attribute called TYPE, which is defined to have a different set of possible values for each tag element. The markup is described in section 2, below.

Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition

Erik F. Tjong Kim Sang and **Fien De Meulder**
CNTS - Language Technology Group
University of Antwerp
{erikt,fien.demeulder}@uia.ua.ac.be

Abstract

We describe the CoNLL-2003 shared task: language-independent named entity recognition. We give background information on the data sets (English and German) and the evaluation method, present a general overview of the systems that have taken part in the task and discuss their performance.

1 Introduction

Named entities are phrases that contain the names of persons, organizations and locations. Example:

> [ORG U.N.] official [PER Ekeus] heads for [LOC Baghdad] .

This sentence contains three named entities: *Ekeus* is a person, *U.N.* is a organization and *Baghdad* is a location. Named entity recognition is an important task of information extraction systems. There has been a lot of work on named entity recognition, especially for English (see Borthwick (1999) for an overview of the systems that have taken part in the 2003 shared task have been offered training and test data for two other European languages: English and German. They have used the data for developing a named-entity recognition system that includes a machine learning component. The shared task organizers were especially interested in approaches that made use of resources other than the supplied training data, for example gazetteers and unannotated data.

2 Data and Evaluation

In this section we discuss the sources of the data that were used in this shared task, the preprocessing steps we have performed on the data, the format of the data and the method that was used for evaluating the participating systems.

2.1 Data

The CoNLL-2003 named entity data consists of eight files covering two languages: English and German[1]. For each of the languages there is a training file, a development file, a test file and a large file with unannotated data. The learning methods were trained with the training data. The development data could be

Annotation Guidelines for

Entity Detection and Tracking (EDT)

Version 4.2.6 200400401

1 Introduction

The Entity Detection task requires that selected types of entities mentioned in the source data be detected, their sense disambiguated, and that selected attributes of these entities be extracted and merged into a unified representation for each entity.

Basic Concepts

- An **entity** is an object or set of objects in the world.
- A **mention** is a textual reference to an entity.

Entities may be referenced in a text by their name, indicated by a common noun or noun phrase, or represented by a pronoun. For example, the following are several mentions of a single entity:

> ***Name Mention:*** *Joe Smith*
> ***Nominal Mention:*** *the guy wearing a blue shirt*
> ***Pronoun Mentions****: he, him*

For Phase 3 of ACE, entities are limited to the following seven types:

- **Person** - Person entities are limited to humans. A person may be a single individual or a group.
- **Organization** - Organization entities are limited to corporations, agencies, and other groups of people defined by an established organizational structure.
- **Facility** - Facility entities are limited to buildings and other permanent man-made structures and real estate improvements.
- **Location** - Location entities are limited to geographical entities such as geographical areas and landmasses, bodies of water, and geological

Entity types

- MUC-7: Person, Organization, Location
- CoNLL-2003: Person (PER), Organisation (ORG), Location (LOC), Miscellaneous (MISC)
- ACE 2004: Person (PER), Organization (ORG), Location (LOC), Facility (FAC), Geo-Political Entity (GPE), Vehicles (VEH), Weapon (WEA)

Entity types

- MUC-7: Person, Organization, Location
- CoNLL-2003: Person (PER), Organisation (ORG), Location (LOC), Miscellaneous (MISC)
- ACE 2004: Person (PER), Organization (ORG), Location (LOC), Facility (FAC), Geo-Political Entity (GPE), Vehicles (VEH), Weapon (WEA)

$\hookrightarrow$ + Date, Product, Event, Money...

Entity types: OntoNotes 5.0

PERSON	People, including fictional
NORP	Nationalities or religious or political groups
FACILITY	Buildings, airports, highways, bridges, etc.
ORGANIZATION	Companies, agencies, institutions, etc.
GPE	Countries, cities, states
LOCATION	Non-GPE locations, mountain ranges, bodies of water
PRODUCT	Vehicles, weapons, foods, etc. (Not services)
EVENT	Named hurricanes, battles, wars, sports events, etc.
WORK OF ART	Titles of books, songs, etc.
LAW	Named documents made into laws
DATE	Absolute or relative dates or periods
TIME	Times smaller than a day
PERCENT	Percentage (including "%")
MONEY	Monetary values, including unit
QUANTITY	Measurements, as of weight or distance
ORDINAL	"first", "second"
CARDINAL	Numerals that do not fall under another type

Encoding scheme for sequence labelling

U.N.	NNP	I-NP	I-ORG
official	NN	I-NP	O
Ekeus	NNP	I-NP	I-PER
heads	VBZ	I-VP	O
for	IN	I-PP	O
Baghdad	NNP	I-NP	I-LOC
.	.	O	O

BIO & variants

- BIO: Begin, Inside, Outside
- IOB: Inside, Outside, Begin (if ambiguous)
- BILUO (=BILOU): Begin, Inside, Last, Unique, Outside
- BIOES: Begin, Inside, Outside, End, Single
- IOBES...

BIO variants

	John	D.	Smith	arrived	in	Toronto	yesterday
BIO	B-PER	I-PER	I-PER	O	O	B-LOC	B-DATE
IOB	I-PER	I-PER	I-PER	O	O	I-LOC	I-DATE
BILUO	B-PER	I-PER	L-PER	O	O	U-LOC	U-DATE
BIOES	B-PER	I-PER	E-PER	O	O	S-LOC	S-DATE

“*Marc graduated from Université Paris Cité*”

“*Marc graduated from Université Paris Cité*”

Flat NER versus Nested NER

Formats for nested NER: SGML

Inline:

The <ENAMEX TYPE="LOCATION">U.K.</ENAMEX> satellite television broadcaster said its subscriber base grew <NUMEX TYPE="PERCENT">17.5 percent</NUMEX> during <TIMEX TYPE="DATE">the past year</TIMEX> to 5.35 million

Standoff:

```
−<source_file URI="VOA20001231.2000.3520.sgm" SOURCE="broadcast news" TYPE="text" VERSION="4.0" AUTHOR="LDC" ENCODING="UTF-8">
  −<document DOCID="VOA20001231.2000.3520">
    −<entity ID="VOA20001231.2000.3520-E1" TYPE="PER" CLASS="SPC">
      −<entity_mention ID="1-2" TYPE="PRO" LDCTYPE="PTV">
        −<extent>
            <charseq START="249" END="265">most of the world</charseq>
          </extent>
        −<head>
            <charseq START="249" END="252">most</charseq>
          </head>
        </entity_mention>
      </entity>
    −<entity ID="VOA20001231.2000.3520-E2" TYPE="PER" CLASS="SPC">
      −<entity_mention ID="2-1" TYPE="NAM" LDCTYPE="NAM" LDCATR="FALSE">
        −<extent>
            <charseq START="65" END="89">Cuban leader Fidel Castro</charseq>
          </extent>
        −<head>
            <charseq START="78" END="89">Fidel Castro</charseq>
          </head>
        </entity_mention>
      −<entity_mention ID="2-3" TYPE="PRE" LDCTYPE="PRE" LDCATR="TRUE">
        −<extent>
            <charseq START="65" END="76">Cuban leader</charseq>
```

Formats for nested NER: brat standoff

Text files (`.txt`)

Text files are expected to have the suffix `.txt` and contain the text of the original documents input into the system.

```
Sony formed a joint venture with Ericsson, a mobile phone
company based in Sweden.
Sony announced today that ...
```

The document texts are stored in plain text files encoded using UTF-8 (an extension of ASCII — plain ASCII texts work also). Document texts may contain newlines, which will be shown as line breaks by brat. However, it is not necessary for the documents to contain any newlines: brat can perform its own sentence segmentation for display using a reliable algorithm. (Whether or not newlines are included in the original text documents, the text files themselves are not modified.)

Annotation files (`.ann`)

Annotations are stored in files with the `.ann` suffix. The various annotation types that may be contained in these files are discussed in the following.

General annotation structure

All annotations follow the same basic structure: Each line contains one annotation, and each annotation is given an ID that appears first on the line, separated from the rest of the annotation by a single TAB character. The rest of the structure varies by annotation type.

Examples of annotation for an entity (`T1`), an event trigger (`T2`), an event (`E1`) and a relation (`R1`) are shown in the following.

```
T1      Organization 0 4                    Sony
T2      MERGE-ORG 14 27                     joint venture
T3      Organization 33 41                  Ericsson
E1      MERGE-ORG:T2 Org1:T1 Org2:T3
T4      Country 75 81                       Sweden
R1      Origin Arg1:T3 Arg2:T4
```

NER corpora

- CoNLL 2003 (English & German, news) $\rightarrow$ $\approx$20k sentences per language
- ACE 2004 & ACE 2005 (nested, English / Chinese / Arabic, multi-genre)
- Genia (nested, English, biomedical)
- OntoNotes (nested, English / Chinese / Arabic, multi-genre)
- WikiNER (automatic labelling, 9 languages, Wikipedia articles) $\rightarrow$ $\approx$200k sentences per language
- ...

NER corpora for French

- ESTER (news transcripts)
- Quaero (nested)
- WiNER-fr (nested)
- French TreeBank
- WikiNER (automatic labelling)
- ...

Methods for NER

Methods for NER

- Regex: `/[A-Z][a-z]* [A-Z][a-z]*/`, `/\d\d-\d\d-\d\d\d\d/`, `/\W+, Inc./`

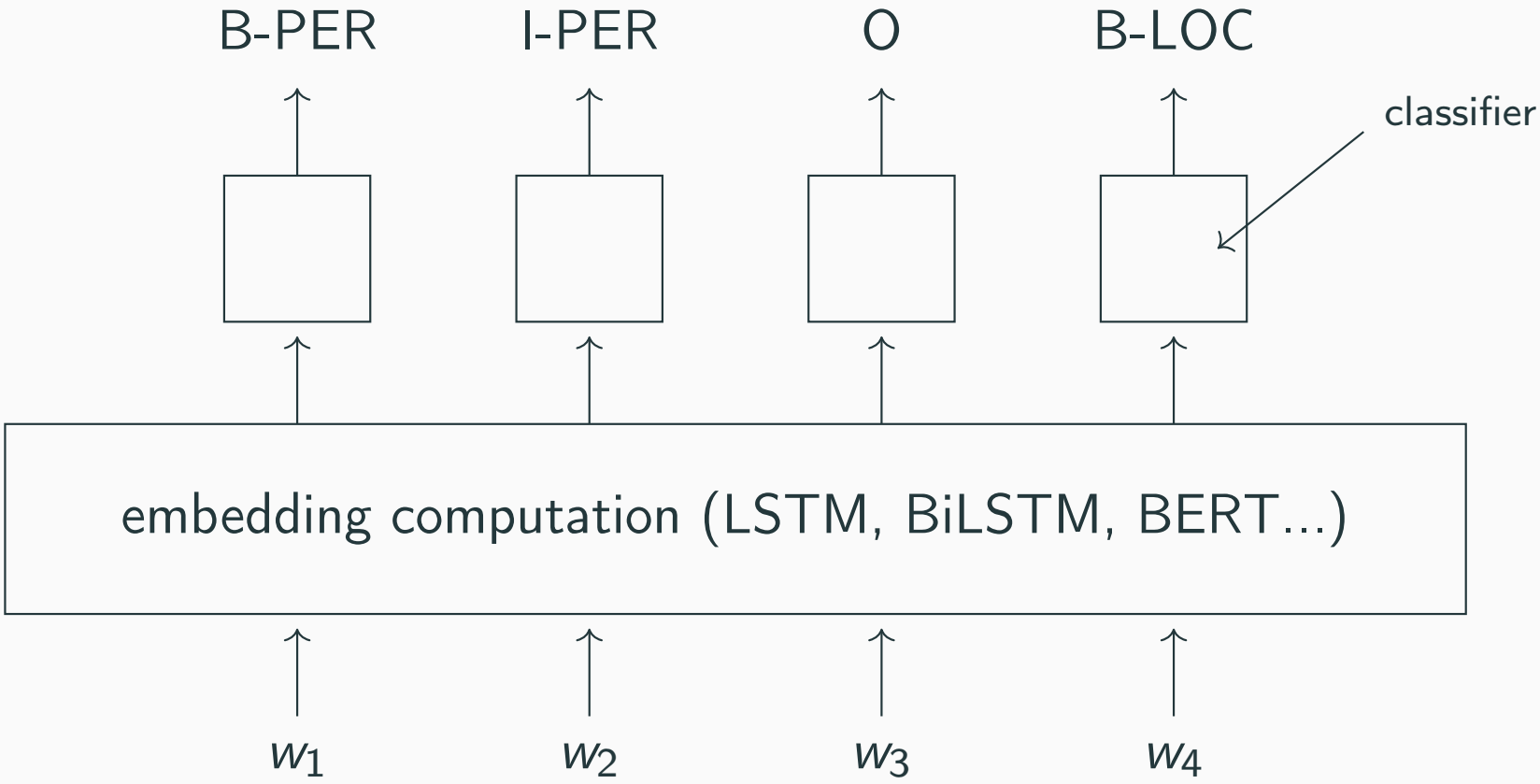
Methods for NER

- Regex: `/[A-Z][a-z]* [A-Z][a-z]*/`, `/\d\d-\d\d-\d\d\d\d/`, `/\W+, Inc./`
- PoS tagging $\rightsquigarrow$ feature `PROPN`

Methods for NER

- Regex: `/[A-Z][a-z]* [A-Z][a-z]*/`, `/\d\d-\d\d-\d\d\d\d/`, `/\W+, Inc./`
- PoS tagging $\rightsquigarrow$ feature `PROPN`
- Context: *Gujibn Poehi says hi*
- ...

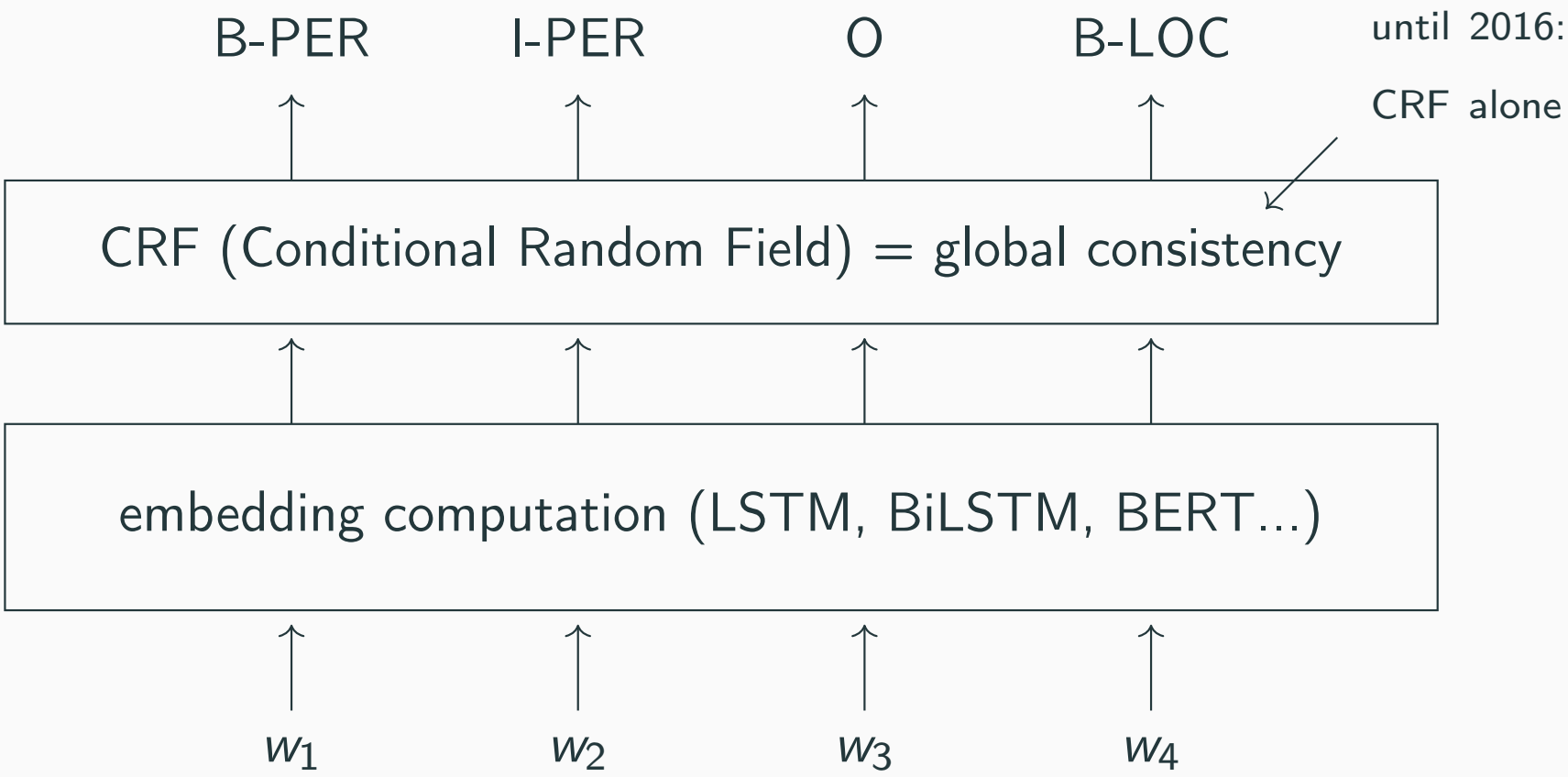
Recent methods for NER



$\rightsquigarrow$ **No more need for gazetteers**

... but still useful when they exist? e.g. Wikipedia

Recent methods for NER



$\rightsquigarrow$ **No more need for gazetteers**

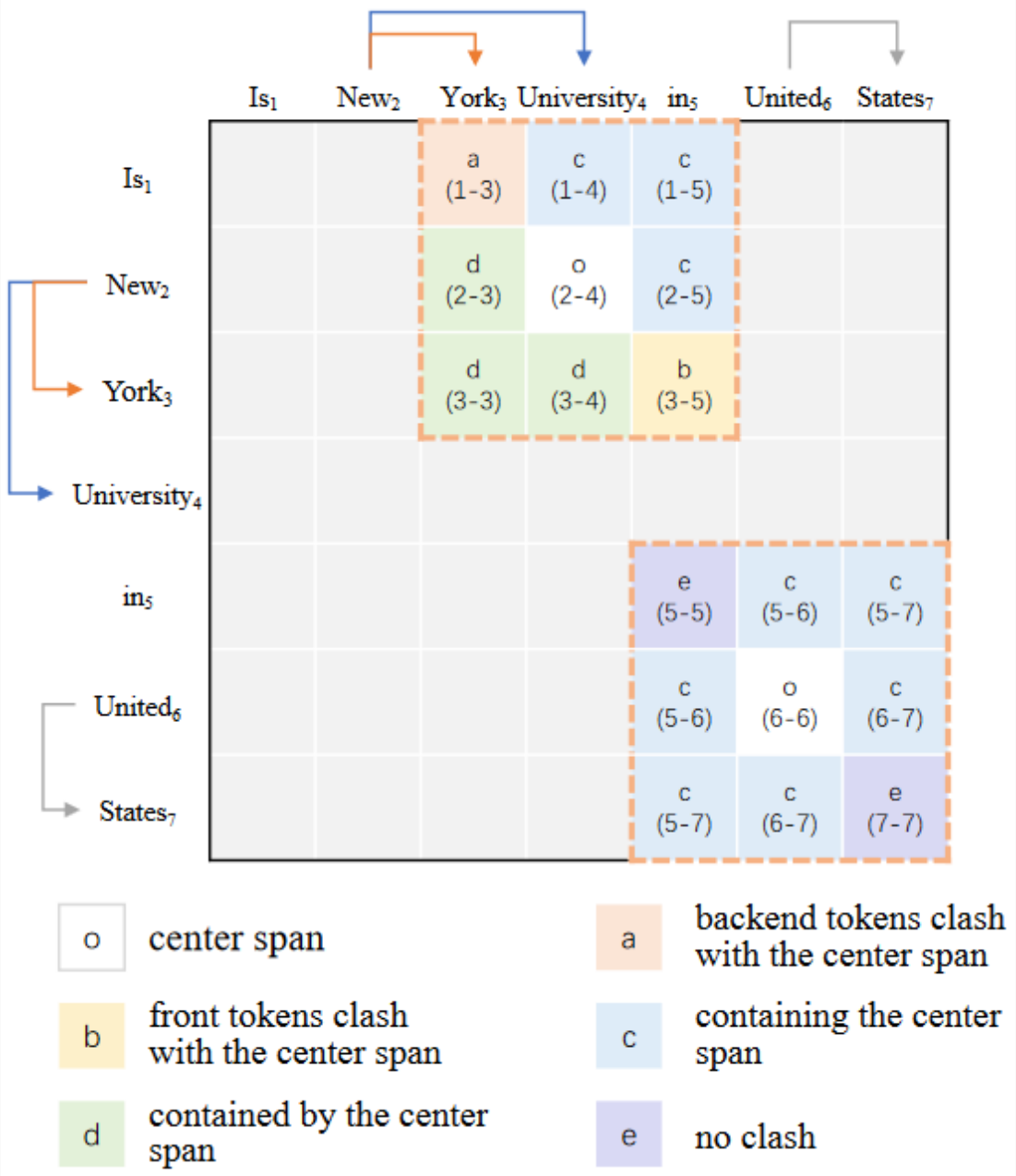
... but still useful when they exist? e.g. Wikipedia

Another method: shift-reduce

Transition	Output	Stack	Buffer	Segment
	[]	[]	[Mark, Watney, visited, Mars]	
SHIFT	[]	[Mark]	[Watney, visited, Mars]	
SHIFT	[]	[Mark, Watney]	[visited, Mars]	
REDUCE(PER)	[(Mark Watney)-PER]	[]	[visited, Mars]	(Mark Watney)-PER
OUT	[(Mark Watney)-PER, visited]	[]	[Mars]	
SHIFT	[(Mark Watney)-PER, visited]	[Mars]	[]	
REDUCE(LOC)	[(Mark Watney)-PER, visited, (Mars)-LOC]	[]	[]	(Mars)-LOC

Lample et al., 2016

Yet another method: CNN-NER



Yan et al., 2023

Pretrained tools for NER

- **Stanford CoreNLP:** Chinese, English, French, German, Hungarian, Italian, Spanish
- **spaCy:** Catalan, Chinese, Croatian, Danish, Dutch, English, Finnish, French, German, Greek, Italian, Japanese, Korean, Lithuanian, Macedonian, Norwegian Bokmål, Polish, Portuguese, Romanian, Russian, Slovenian, Spanish, Swedish, Ukrainian
- **Stanza:** Afrikaans, Arabic, Armenian, Bulgarian, Chinese, Danish, Dutch, English, Finnish, French, German, Hungarian, Italian, Japanese, Kazakh, Marathi, Myanmar, Norwegian-Bokmaal, Norwegian-Nynorsk, Persian, Polish, Russian, Sindhi, Spanish, Swedish, Thai, Turkish, Ukrainian, Vietnamese
- ...

Evaluation metrics: precision, recall, F1-score

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

$$F1 = \frac{2 \cdot P \cdot R}{P + R}$$

Applying F1 to NER

Reference:	B-PER	I-PER	O	B-LOC	O
Prediction:	B-PER	O	O	B-LOC	I-LOC

Applying F1 to NER

Reference:	B-PER	I-PER	O	B-LOC	O
Prediction:	B-PER	O	O	B-LOC	I-LOC
	B-PER	I-LOC	O	I-PER	O

Applying F1 to NER

Reference:	B-PER	I-PER	O	B-LOC	O
Prediction:	B-PER	O	O	B-LOC	I-LOC
	B-PER	I-LOC	O	I-PER	O
	B-PER	B-PER	O	B-LOC	O

Micro-F1 and macro-F1

- Micro-average: compute **all** TP / FP / FN, then the metric
- Macro-average: compute TP / FP / FN **for each class**, then the metrics, then average

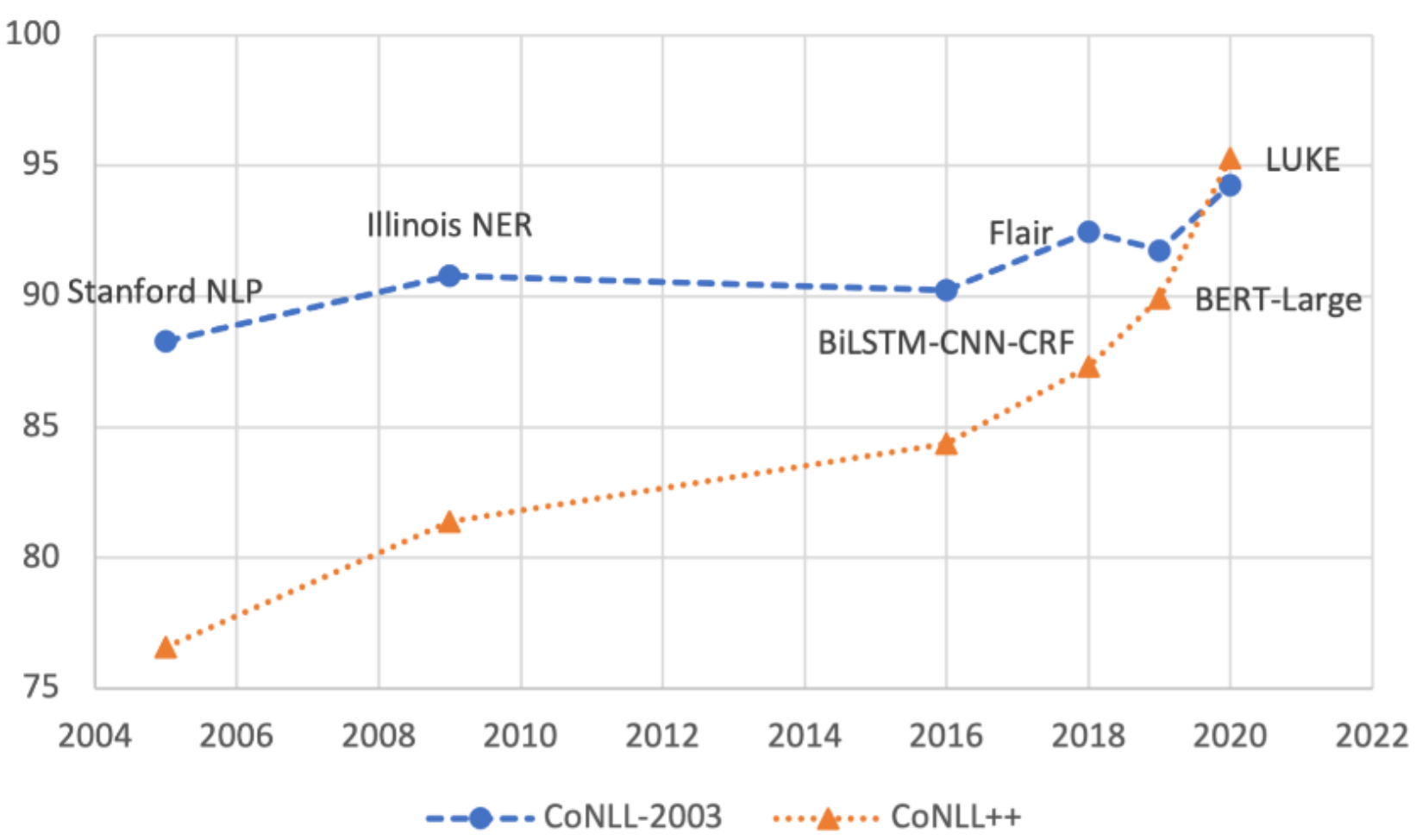
	PER	ORG	LOCATION	macro
Precision	97.5	90.0	81.7	89.7
Recall	95.4	85.3	82.9	87.9
F1	96.4	87.6	82.3	88.8

$\hookrightarrow$ Usually ignoring "MISC" or "Other" classes

Match criteria for NER F1

- **Strict:** Both the boundaries and the entity type must be correct.
- **Boundaries:** Entity type is not considered and boundaries must be correct.
- **Relaxed:** A multi-token entity is considered correct if at least one token is correctly typed (boundaries are ignored).

↪ Always be specific!



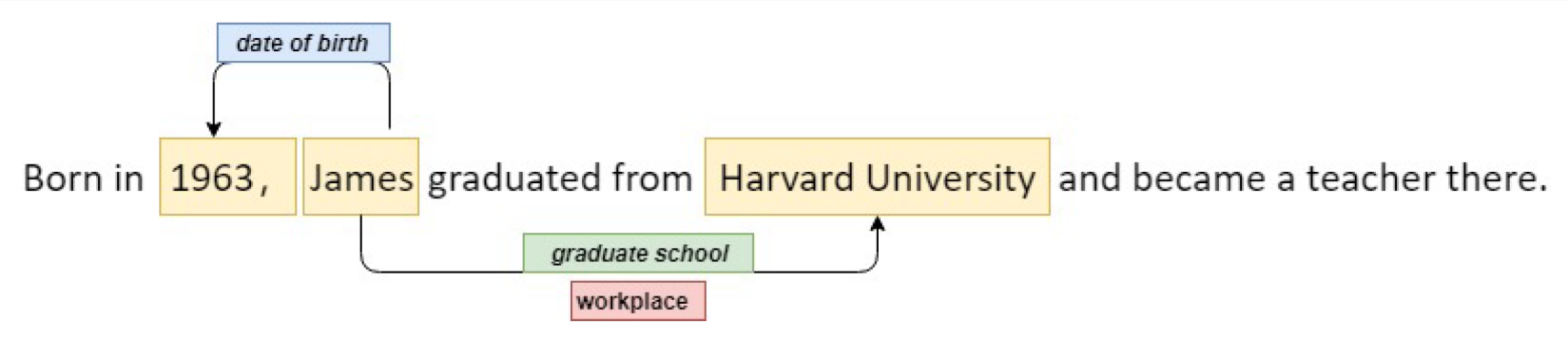
Liu & Ritter, 2023

Relation extraction

"Born in 1963, James graduated from Harvard University and became a teacher there."

Peng & Chen, 2020

"Born in 1963, James graduated from Harvard University and became a teacher there."



Peng & Chen, 2020

Relation Extraction (RE)

Classifying pairs of entities according to a predefined set of relations

Corpora

- CoNLL 2004
- ACE 2004, ACE 2005
- NYT (automatic labelling)
- TACRED
- DocRED (document-level)
- KBP37
- FewRel (few-shot)
- Semantic relation extraction corpora...

Relation types: ACE

Table 1 ACE05 Entity Types and Subtypes

Type	Subtypes
FAC (Facility)	Airport, Building-Grounds, Path, Plant, Subarea-Facility
GPE (Geo-Political Entity[3])	Continent, County-or-District, GPE-Cluster, Nation, Population-Center, Special, State-or-Province
LOC (Location)	Address, Boundary, Celestial, Land-Region-Natural, Region-General, Region-International, Water-Body
ORG (Organization)	Commercial, Educational, Entertainment, Government, Media, Medical-Science, Non-Governmental, Religious, Sports
PER (Person)	Group, Indeterminate, Individual
VEH (Vehicle)	Air, Land, Subarea-Vehicle, Underspecified, Water
WEA (Weapon)	Biological, Blunt, Chemical, Exploding, Nuclear, Projectile, Sharp, Shooting, Underspecified

Table 6 ACE05 Relation Types and Subtypes
(Relations marked with an * are symmetric relations.)

Type	Subtype
ART (artifact)	User-Owner-Inventor-Manufacturer
GEN-AFF (Gen-affiliation)	Citizen-Resident-Religion-Ethnicity, Org-Location
METONYMY*	*none*
ORG-AFF (Org-affiliation)	Employment, Founder, Ownership, Student-Alum, Sports-Affiliation, Investor-Shareholder, Membership
PART-WHOLE (part-whole)	Artifact, Geographical, Subsidiary
PER-SOC* (person-social)	Business, Family, Lasting-Personal
PHYS* (physical)	Located, Near

From Yao & Zhong

Relation types: SemEval-2010 Task 8

Content-Container (CC). An object is physically stored in a delineated area of space. Example: *a bottle full of honey was weighed*

Entity-Origin (EO). An entity is coming or is derived from an origin (e.g., position or material). Example: *letters from foreign countries*

Entity-Destination (ED). An entity is moving towards a destination. Example: *the boy went to bed*

Component-Whole (CW). An object is a component of a larger whole. Example: *my apartment has a large kitchen*

Member-Collection (MC). A member forms a nonfunctional part of a collection. Example: *there are many trees in the forest*

Message-Topic (MT). A message, written or spoken, is about a topic. Example: *the lecture was about semantics*

Cause-Effect (CE). An event or object leads to an effect. Example: *those cancers were caused by radiation exposures*

Instrument-Agency (IA). An agent uses an instrument. Example: *phone operator*

Product-Producer (PP). A producer causes a product to exist. Example: *a factory manufactures suits*

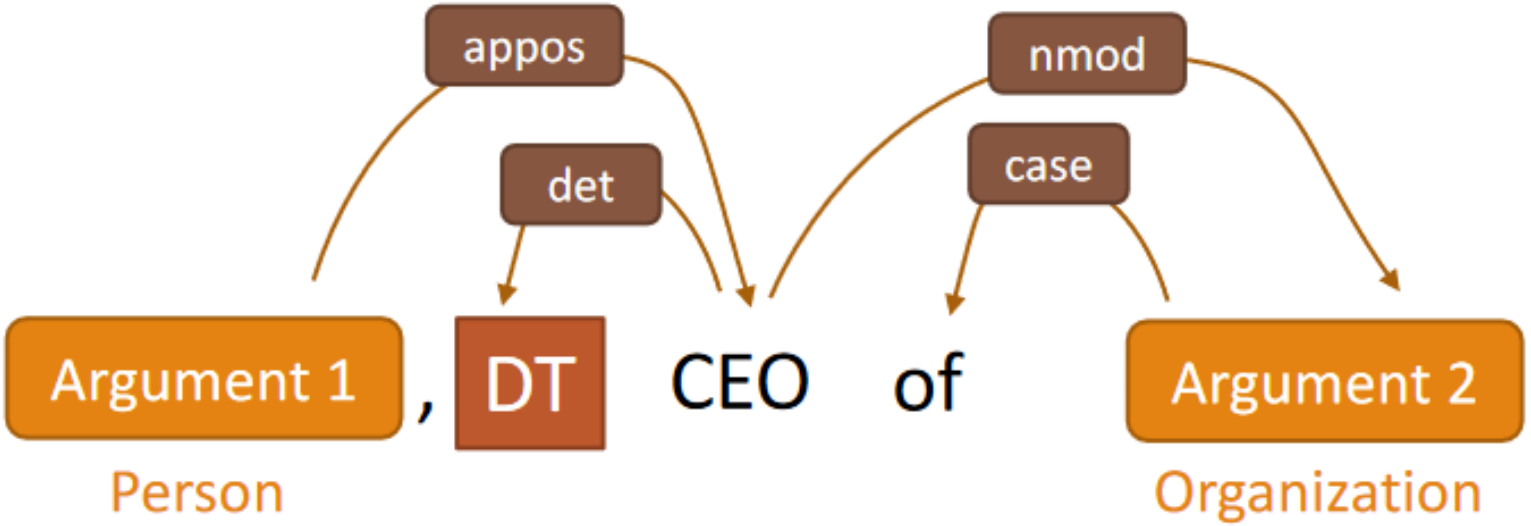
From Yao & Zhong

Hearst patterns (1992)

Generic grammatical patterns for the **is-a** relation:

- X and other Y
- X or other Y
- Y such as X
- Such Y as X
- Y including X
- Y , especially Y

Complex lexico-syntactic patterns



Bill Gates, the CEO of Microsoft, said ...
Mr. Jobs, the brilliant and charming CEO of Apple Inc., said ...
... announced by Steve Jobs, the CEO of Apple.
... announced by Bill Gates, the director and CEO of Microsoft.
... mused Bill, a former CEO of Microsoft.
and many other possible instantiations...

$\rightsquigarrow$ And when adding concepts: *lexico-semantic* patterns

From Sameer Singh

Bootstrapping: semi-supervision through patterns

- Start with a seed: high-precision patterns or high-confidence triples

- Iterate:
  - Search a large corpus for sentences containing known entity pairs
  - Extract patterns
  - Filter with heuristics (frequency, precision...)
  - Add to seed patterns and reapply to get new entity pairs

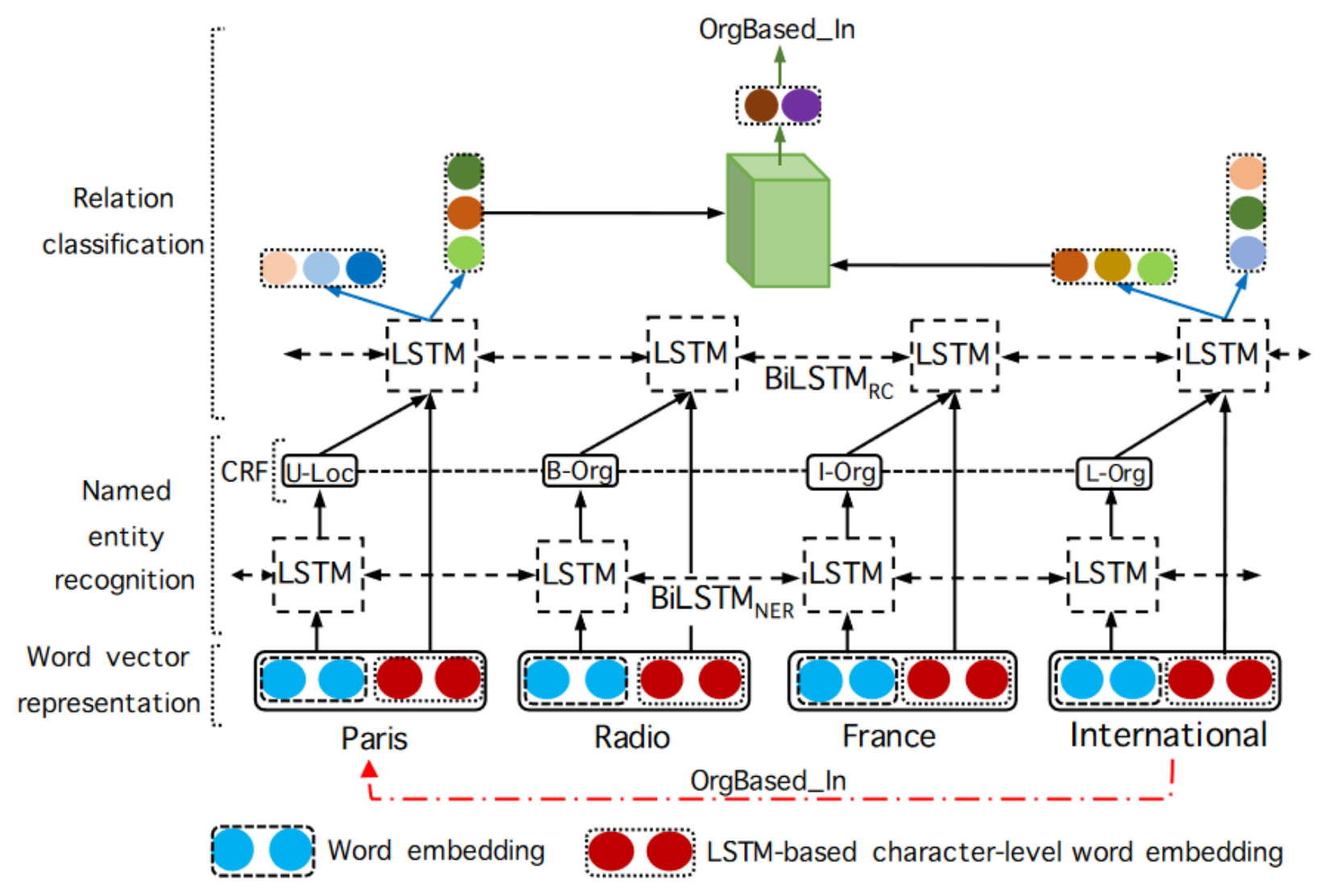
Feature-based ML

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

Entity-based features		
	Entity_1 type	ORG
	Entity_1 head	*airlines*
	Entity_2 type	PERS
	Entity_2 head	*Wagner*
	Concatenated types	ORGPERS
Word-based features		
	Between-entity bag of words	{ *a, unit, of, AMR, Inc., immediately, matched, the, move, spokesman* }
	Word(s) before Entity_1	NONE
	Word(s) after Entity_2	*said*
Syntactic features		
	Constituent path	$NP \uparrow NP \uparrow S \uparrow S \downarrow NP$
	Base syntactic chunk path	$NP \rightarrow NP \rightarrow PP \rightarrow NP \rightarrow VP \rightarrow NP \rightarrow NP$
	Typed-dependency path	$Airlines \leftarrow_{subj} matched \leftarrow_{comp} said \rightarrow_{subj} Wagner$

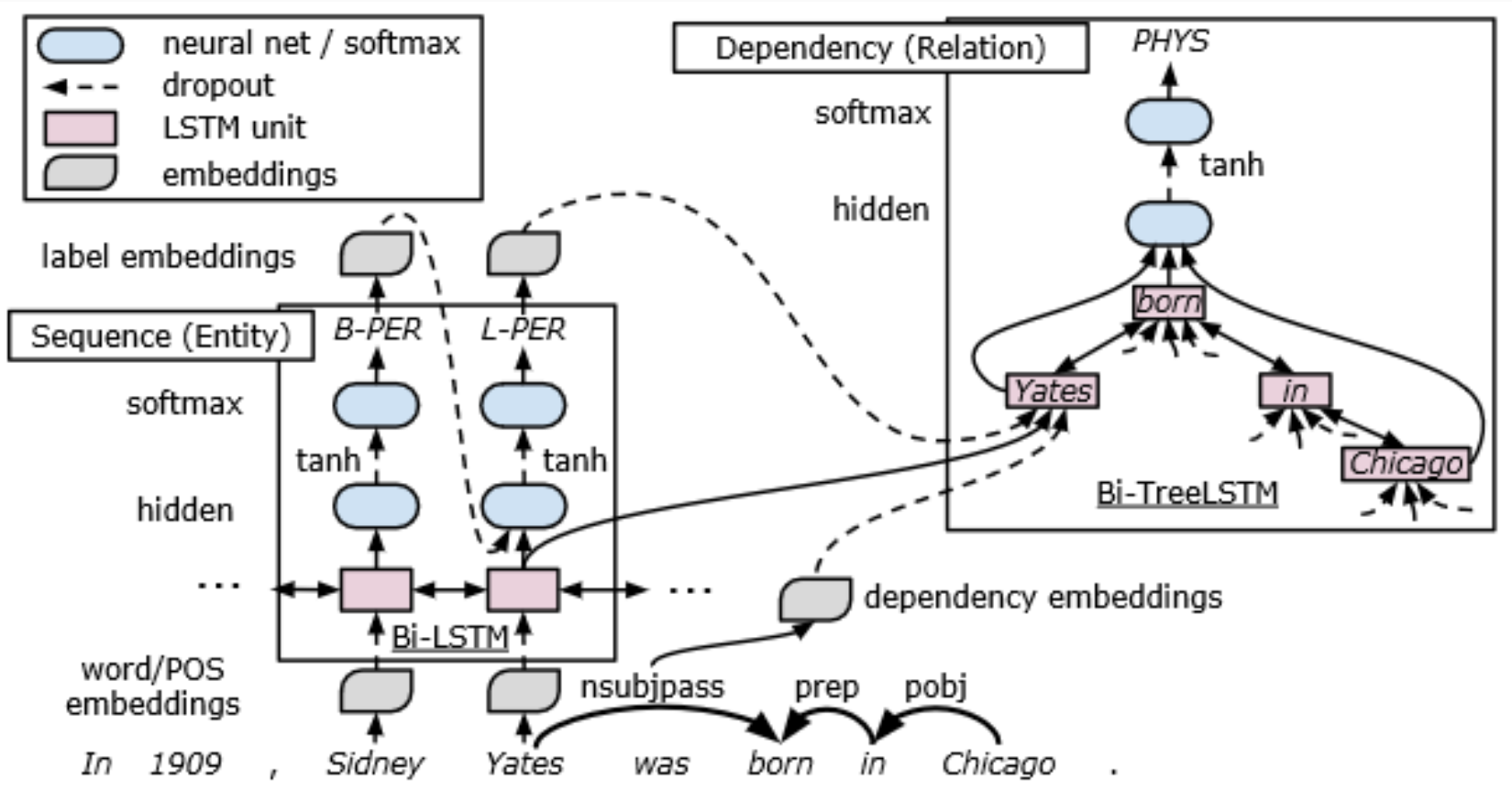
From Sameer Singh

Classifying embedding pairs



Nguyen & Verspoor, 2019

Linguistically-informed neural models

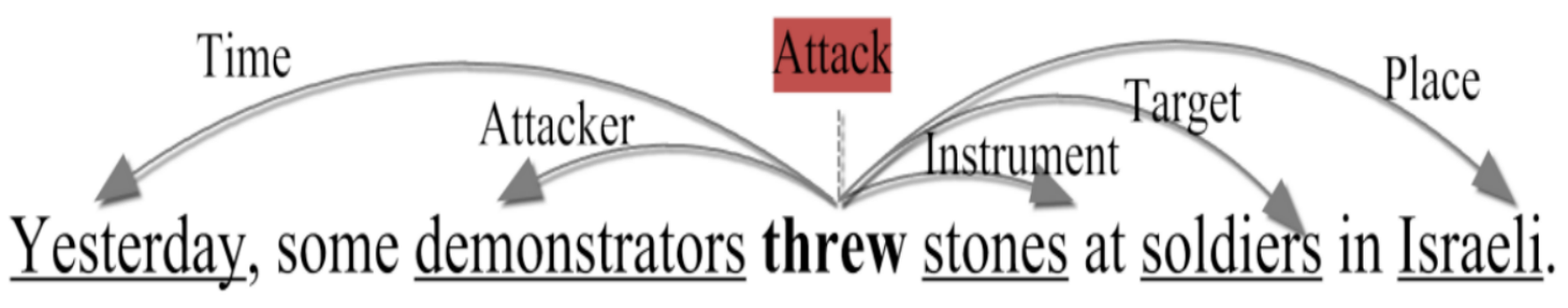


N-ary relation extraction

"Yesterday, some demonstrators threw stones at soldiers in Israeli."

Chen et al., 2017

"Yesterday, some demonstrators threw stones at soldiers in Israeli."



Chen et al., 2017

A versatile formalism

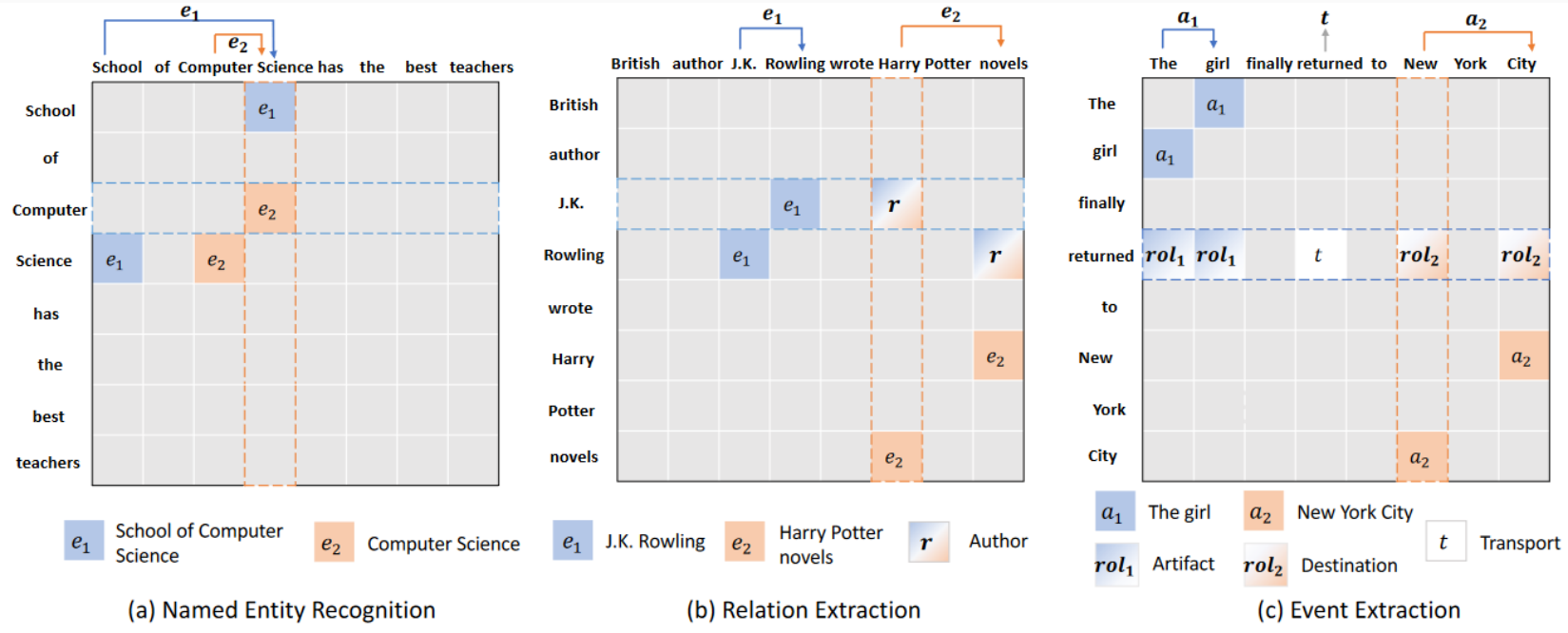
- Trigger
- Arguments
- Roles

$\rightsquigarrow$ Event extraction, frame-semantic parsing (FrameNet)...

Corpora

- ACE 2005
- Genia
- ...

Encoding all 3 tasks with token pairs



Yan et al., 2023

See you next week!

first.last@inria.fr