Automatic Knowledge Extraction from Documents

J. Fan
A. Kalyanpur
D. C. Gondek
D. A. Ferrucci

Julia Chapple
Jack Ma

Introduction

After parsing Watson has a large corpus of data but no knowledge.

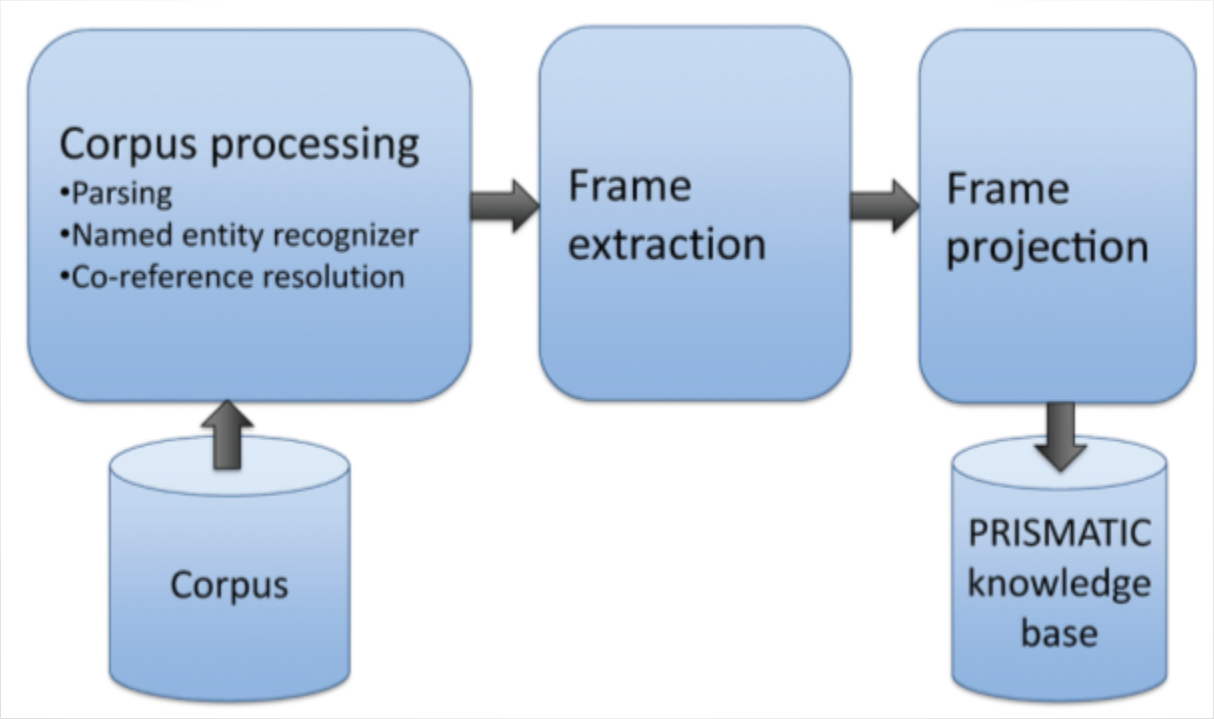
To use the corpus Watson must be able to:

- Exact facts and relationships from corpus
- Search over the corpus to retrieve a relevant fact

PRISMATIC converts the corpus to useful knowledge

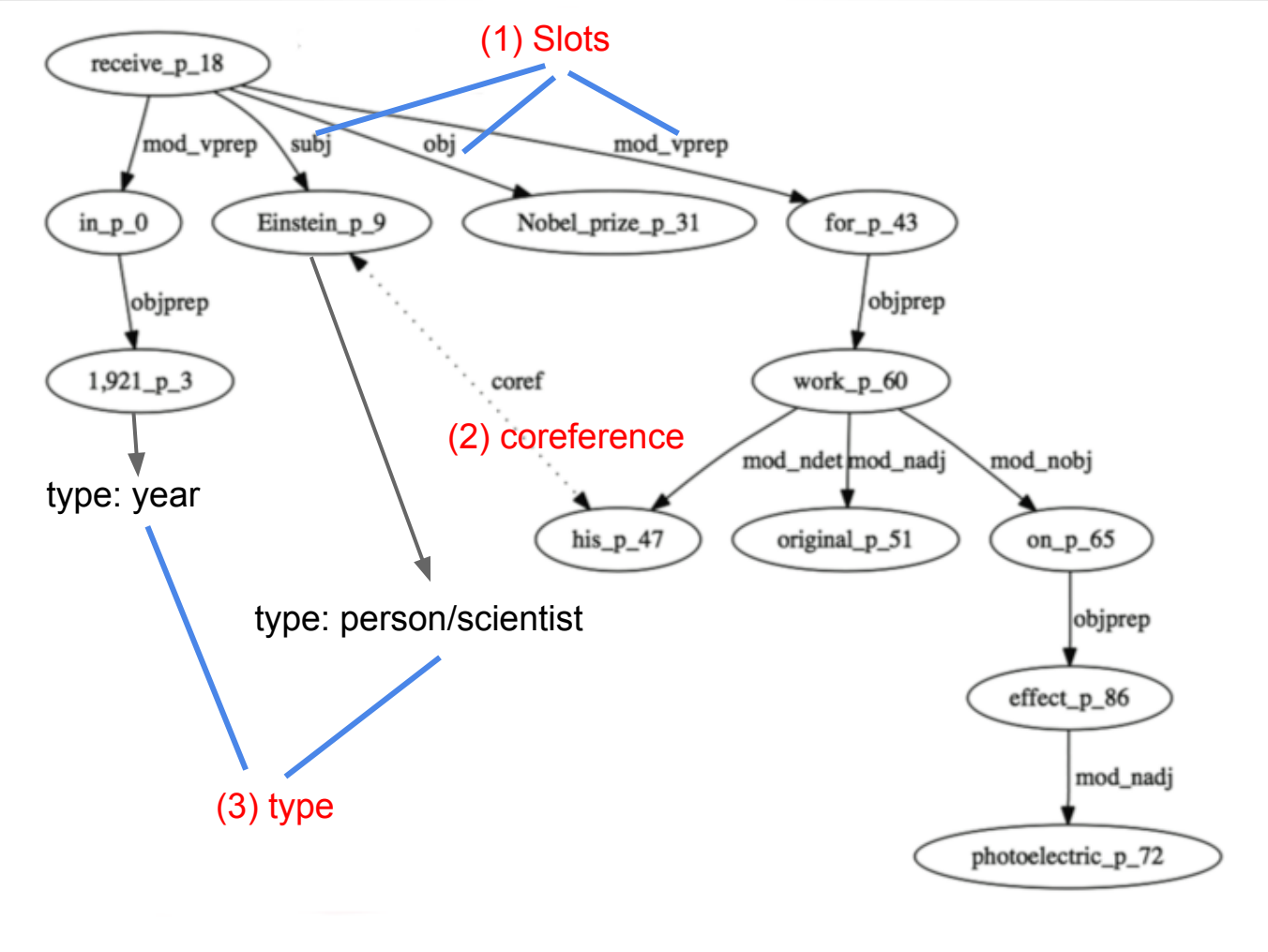
Overview

- Determine the relationships within sections of text.
- Infer facts from aggregate statistics of relationships



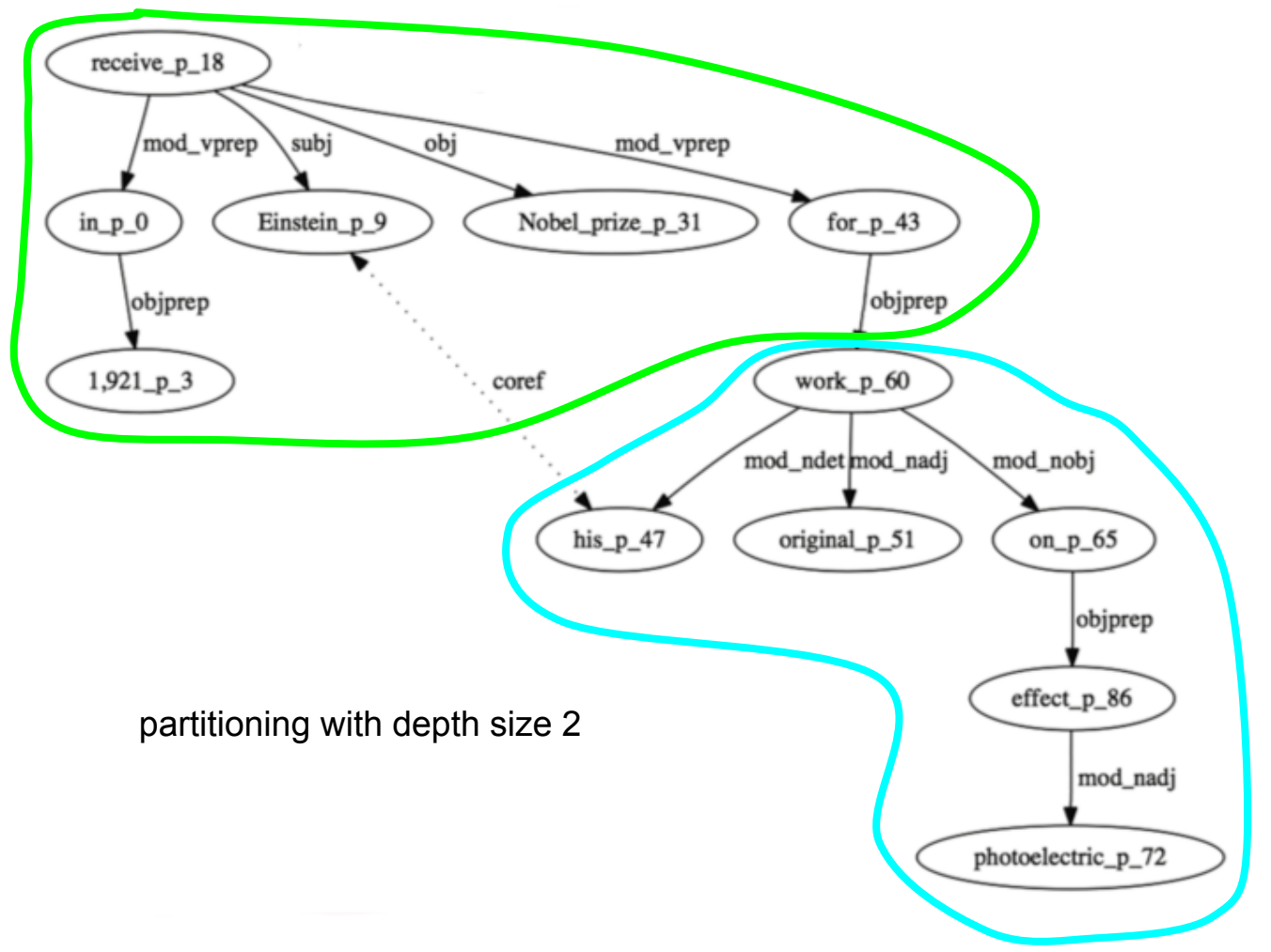
Corpus Processing

In 1921, Einstein received the Nobel Prize for his original work on the photoelectric effect.



Frame Extraction (1/2)

In 1921, Einstein received the Nobel Prize for his original work on the photoelectric effect.



Frame Extraction (2/2)

In 1921, Einstein received the Nobel Prize for

	Frame01
verb	receive
subj	Einstein
type	PERSON/SCIENTIST
obj	Nobel prize
mod_vprep	in
objprep	1921
type	YEAR
mod_vprep	for
objprep	Frame02

his original work on the photoelectric effect.

	Frame02
noun	work
mod_ndet	his/Einstein
mod_nobj	on
objprep	effect

Projection over a Frame

Finding frames that match constraints over certain relations between words.

In 1921, Einstein received the Nobel Prize for Frame02

Einstein receive Nobel prize

{ (verb, "receive")
(subj, "Einstein")
(type, PERSON/SCIENTIST)
(obj, "Nobel prize")
(mod_vprep, "in")
(objprep, "1921")
(type, YEAR)
(mod_vprep "for")
(objprep Frame02) }

Subject-Verb-Object (S-V-O)

→

{(verb, "receive")
(subj, "Einstein")
(obj, "Nobel prize") }

Projection over all frames

PRISMATIC generates projections for all frames in the corpus.

Subject-Verb-Object
(S-V-O)

{ frame1, frame2,..., frame_M} → { proj1, proj2,..., proj_U}

Subject-Verb-Preposition-Object
(S-V-P-O)

{ frame1, frame2,..., frame_M} → { proj1, proj2,..., proj_V}

Noun-Isa
(N-Isa)

..... → .....

.....

..... → .....

Statistics

Used to determine common pieces of information and infer axioms

Frequency

How many times a frame occurs in the corpus

Conditional Probability

The likelihood of a frame given some other slot values of the frame

NPMI

Conditional probability with the popularity of slot values accounted for

Application: Type Coercion

- Checks that candidate answers match the LAT of the question
- Evaluates candidates with aggregate statistics on the projection:

  { *noun* answer, *is-a* LAT }
- Improves accuracy 2.4%

Example: Type Coercion

THE SPACE AGE BEGAN OCTOBER 4, 1957 WITH THE LAUNCH OF THIS *SATELLITE*

LAT: Satellite

Candidates: Rocket
Soviet Union
Cold War
Sputnik

Test:

{ *noun* answer, *is-a* satellite }

Best Fit: Sputnik

Application: Type Inference

- In some questions the LAT is meaningless (e.g. *this* or *it*)

- PRISMATIC attempts to determine the type from lexical relationships
  - Finds the type that fits into the frame of the question
  - Cannot take the context of the question into account

Examples: Type Inference

NATURALLY, *IT'S* NIGER'S NEIGHBOR

Projection:

{ *type* Region, *verb* neighbor, *type* ?? }

Best Fit: Region

IN THE BILLIARDS GAME NAMED FOR THIS BLACK OBJECT, YOU MUST SINK *IT* LAST

Projection:

{ *verb* sink, *type* ?? }

Best Fit: Ship

Application: Candidate Generation

- PRISMATIC can also generate candidate answers
- Uses the LAT and its modifiers to find the 20 most common instances
- One of the better candidate generation subsystems
- Guesses well even when Watson can't understand the question

Example: Candidate Generation

THE SUITS IN THIS *DECK* OF CARDS INCLUDE WANDS, PENTACLES & CUPS

LAT: Deck

Modifier: Cards

Common Instances:

52-Card, Standard, ***Tarot***, Euche, Uno ...

Application: Missing Link Detection

Finding the relating link between the actual answer and other entities within the question.

ON HEARING OF THE DISCOVERY OF GEORGE MALLORY'S BODY, *HE* TOLD REPORTERS *HE* STILL THINKS HE WAS FIRST

Correct Answer:

Edmund Hillary

Projections:

{*subj* George Mallory, *verb* perish, *obj* Mount Everest }

{*subj* Edmund Hillary, *verb* climb, *obj* Mount Everest}

Missing Link: Mount Everest

Conclusion

Prismatic is an important system for formal knowledge representation of corpus data used by other systems of Watson to infer knowledge.

These systems (type coercion, type inference, candidate generation, missing link) will be explored and presented later by others in the class.