Natural Language Processing (10)

Knowledge Acquisition

Daisuke Kawahara

Department of Communications and Computer Engineering,
Waseda University

Lecture Plan

1. Overview of Natural Language Processing
2. Formal Language Theory
3. Word Senses and Embeddings
4. Topic Models
5. Collocations, Language Models, and Recurrent Neural Networks
6. Sequence Labeling and Morphological Analysis
7. Parsing (1)
8. Parsing (2)
9. Transfer Learning
10. Knowledge Acquisition
11. Information Retrieval, Question Answering, and Machine Translation
12. Guest Talk (1): Dr. Chikara Hashimoto (Rakuten Institute of Technology)
13. Guest Talk (2): Dr. Tsubasa Takahashi (LINE Corporation)
14. Project: Survey or Programming (do it yourself)
15. Project Presentation

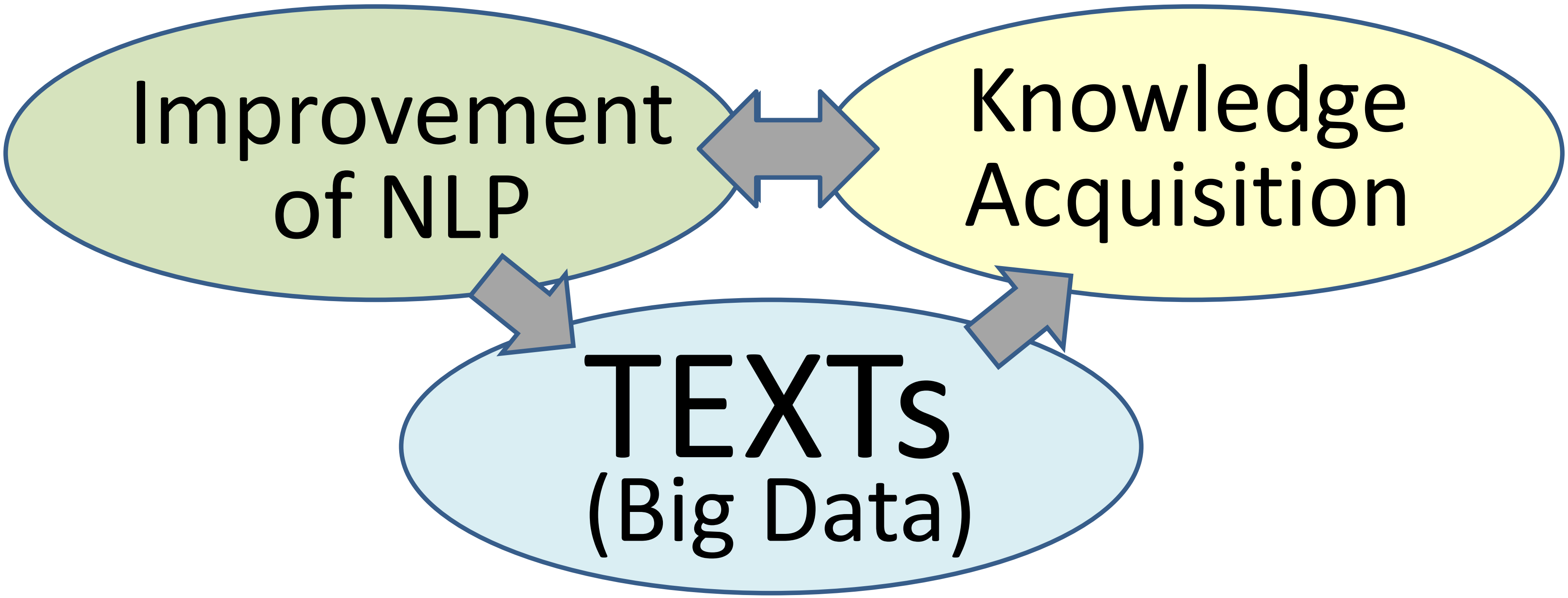

Improvement of NLP
Knowledge Acquisition
TEXTs
(Big Data)

Table of Contents

Knowledge for NLP

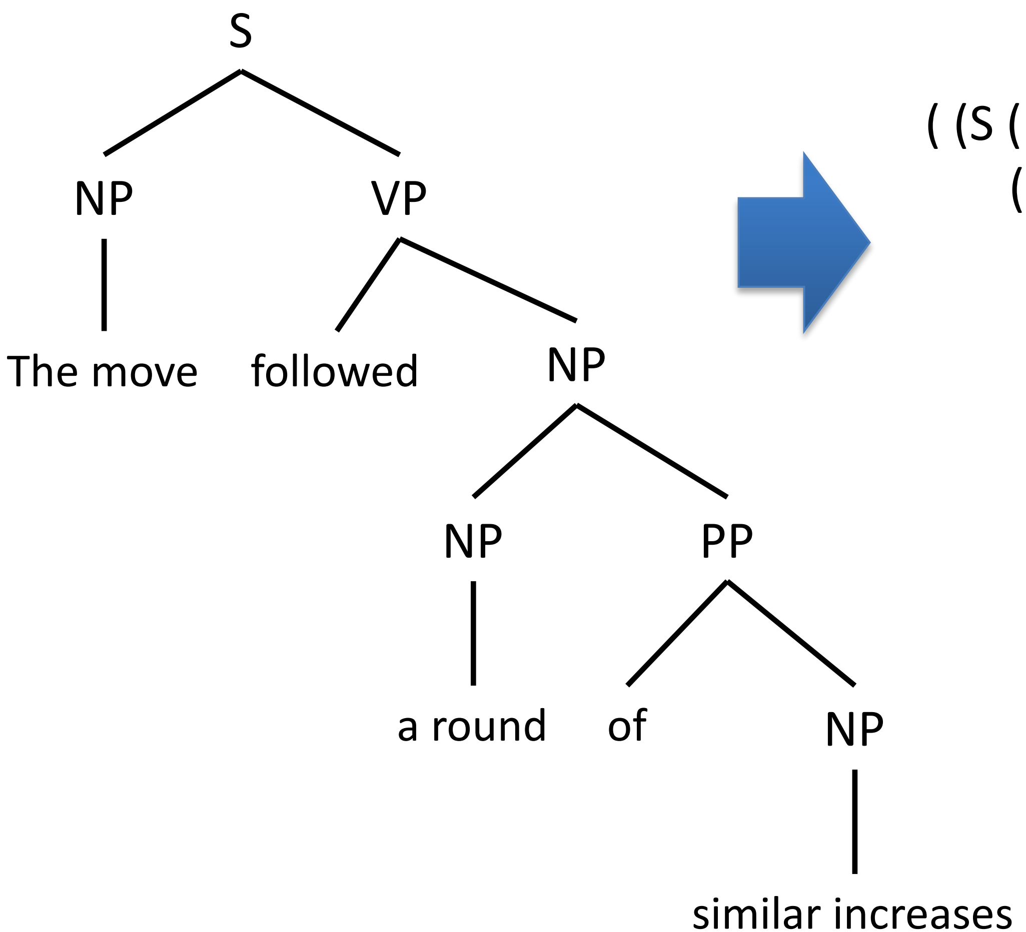
- Grammatical knowledge
  - Vocabulary, part-of-speech (POS), inflection, ...
  - Syntax (how words combine to form a sentence), …
- World knowledge
  - Knowledge between words (phrases): how words (phrases) are semantically related to each other.
  - Knowledge between events: what semantic relation holds between events.

Grammatical Knowledge

- Usually induced from grammatically annotated corpora.
  1. A given text is manually or partly automatically annotated with grammatical information.
  2. Grammatical knowledge is learnt from the annotated text (corpus) manually or automatically.
- With grammatically annotated corpora, computers can learn grammatical knowledge accurately thanks to today's statistical learning methods.
- Grammatically annotated corpora:
  - Penn Treebank (for English)
  - Kyoto University Text Corpus (for Japanese)

Penn Treebank [Marcus+ 1993]

Wall Street Journal (1 million words)




```
( (S (NP The move)
     (VP followed
         (NP (NP a round)
             (PP of
                 (NP similar increases))))))
```

Kyoto University Text Corpus

[Kurohashi and Nagao 1998]

- Mainichi newspaper articles (40K sentences, 1M words)
- Word segmentation, POS, phrase-dependency are annotated

```
# S-ID:950101003-001 KNP:96/10/27 MOD:2005/03/08
* 26D
村山 むらやま 村山 名詞 6 人名 5 * 0 * 0
富市 とみいち 富市 名詞 6 人名 5 * 0 * 0
首相 しゅしょう 首相 名詞 6 普通名詞 1 * 0 * 0
は は は 助詞 9 副助詞 2 * 0 * 0
* 2D
年頭 ねんとう 年頭 名詞 6 普通名詞 1 * 0 * 0
に に に 助詞 9 格助詞 1 * 0 * 0
* 6D
あたり あたり あたる 動詞 2 * 0 子音動詞ラ行 10 基本連用形 8
* 6D
首相 しゅしょう 首相 名詞 6 普通名詞 1 * 0 * 0
官邸 かんてい 官邸 名詞 6 普通名詞 1 * 0 * 0
で で で 助詞 9 格助詞 1 * 0 * 0
* 6D
内閣 ないかく 内閣 名詞 6 普通名詞 1 * 0 * 0
記者 きしゃ 記者 名詞 6 普通名詞 1 * 0 * 0
会 かい 会 名詞 6 普通名詞 1 * 0 * 0
と と と 助詞 9 格助詞 1 * 0 * 0
. . .
```

Knowledge between Words (1/2)

- Synonym (Paraphrase)
  - car = automobile = motorcar = four wheels
  - trade foreign currencies = exchange one currency for another
- IS-A (Hypernym-Hyponym)
  - Barack Obama IS-A politician, politician IS-A human,
  - human IS-A mammal, mammal IS-A living thing, …
- Entailment
  - snore → sleep, gulp → drink, commit apoptosis → die,
  - divorce → marry, get laid off → get hired, …
- Domain
  - EDUCATION: teacher, school, students, textbook, …
  - CULTURE: music, movie, actress, novel, …

Knowledge between Words (2/2)

- Entity set (word class)
  - FRUIT: apple, chokeberry, apricot, cherry, peach, lemon, …
  - ACTOR: Brad Pitt, Leonardo DiCaprio, George Clooney, …
- Other various semantic relations
  - CAUSATION: trauma → PTSD, supercooling → dew condensation, …
  - PREVENTION: encrypt software → information leak, firewall → unauthorized access, …
- Case frames
  - { AOL, M&FC, … } acquire { Time Warner Inc., Nihon Seimitsu Co., Ltd. … }
  - { Shakespeare, Haruki Murakami, … } write { Hamlet, 1Q84, … }

Table of Contents

- Knowledge for NLP
- Case frame acquisition
- Paraphrase acquisition
- Relation extraction
- Entailment acquisition

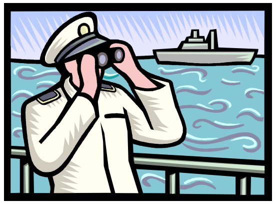
Predicate-Argument Structure

?

<u>クロールで</u> 泳いでいる 女の子を 見た
crawl swim girl saw

?

<u>望遠鏡で</u> 泳いでいる 女の子を 見た
telescope swim girl saw



Case frame

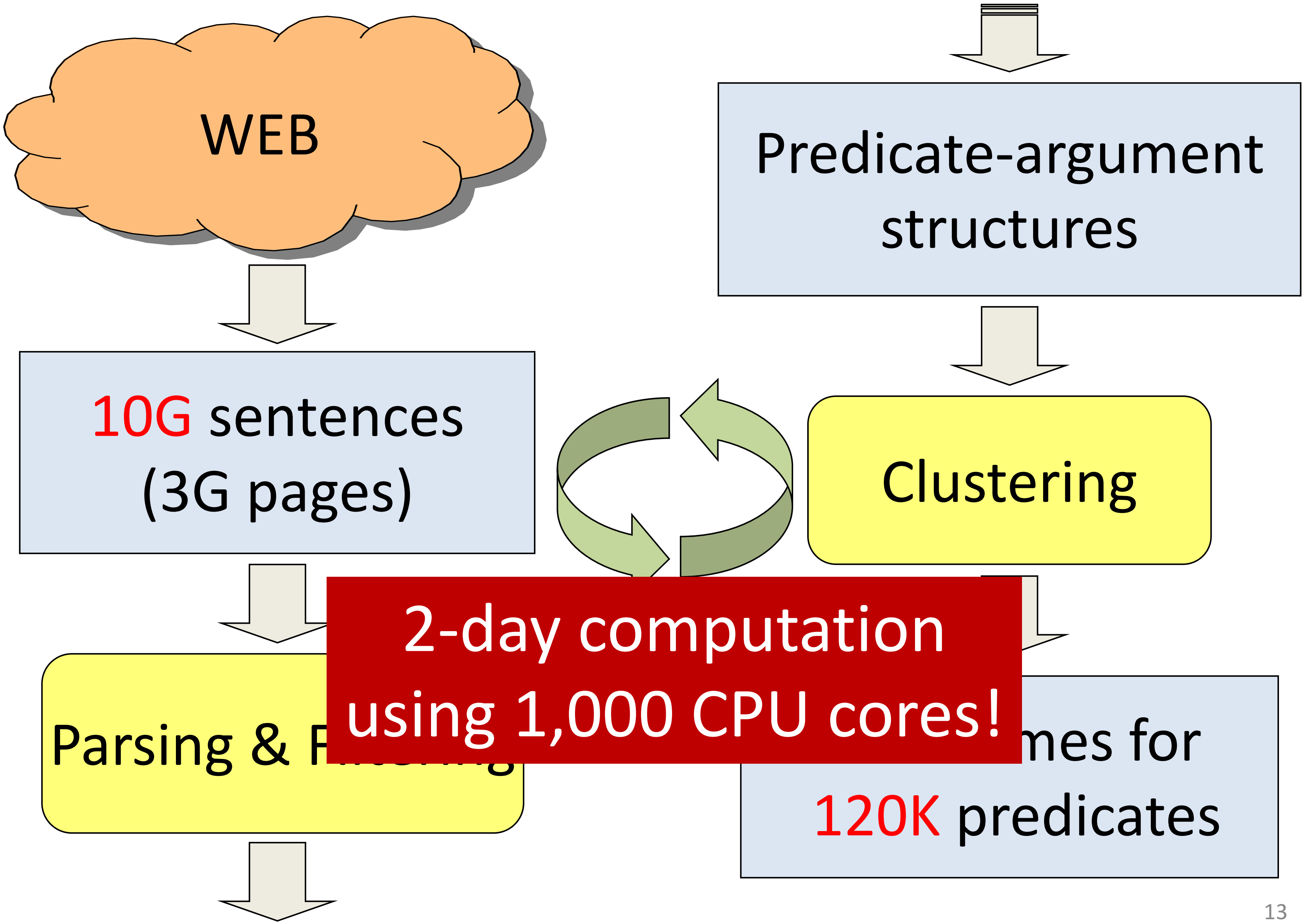
泳ぐ swim

{人 person, 子 child,…}が
{クロール crawl, 平泳ぎ,…}で
{海 sea, 大海,…}を

見る see

{人 person, 者,…}が
{望遠鏡 telescope, 双眼鏡 ,,…}で
{姿 figure, 人 person,…}を

[Kawahara and Kurohashi 2006]



Inducing Semantic Frames and Verb Classes [Kawahara+ 2014]

Verb classes:

investigate

sight

Semantic frames:

observe:1
{they, he, …} observe
{effect, result, …}

observe:2
{you, we, …} observe
{child, people, …}

observe:3
{we, child, …} observe
{bird, wildlife, …}

…

watch:5
{I, we, …} watch
{bird, …}

Verb uses:

they observed the effects
children observed birds
He observed the result
…
…

you observe your child
we observed nice birds
we observed 110 people
…
…

we watch our birds
I watched the movie
We will watch the game
…
…

- The doctor <u>observed</u> the effects of …
- This statistical ability to <u>observe</u> an effect …
- They did not <u>observe</u> a residual effect of …
- We could <u>observe</u> the results at the same time …
- My first opportunity to <u>observe</u> the results of …
- You can <u>observe</u> beautiful birds …
- children may then <u>observe</u> birds …
- …

- The doctor observed the **effects** of …
- This statistical ability to observe an **effect** …
- They did not observe a residual **effect** of …
- We could observe the **results** at the same time …
- My first opportunity to observe the **results** of …
- You can observe beautiful **birds** …
- children may then observe **birds** …
- …

nsubj:{they, doctor, ..} <u>observe</u> dobj:{effect}

nsubj.they:825

nsubj.doctor:235

dobj.effect:5,070

prep_at:{time, point, ..}

prep_at.time:71

prep_at.point:20

nsubj:{doctor, we, ..} <u>observe</u> dobj:{result}

nsubj.doctor:531

nsubj.we:291

dobj.result:2,320

prep_at:{time, point, ..}

prep_at.time:93

prep_at.point:37

nsubj:{you, child, ..} <u>observe</u> dobj:{bird}

nsubj.you:704

nsubj.child:563

dobj.bird:1,692

Chinese Restaurant Process

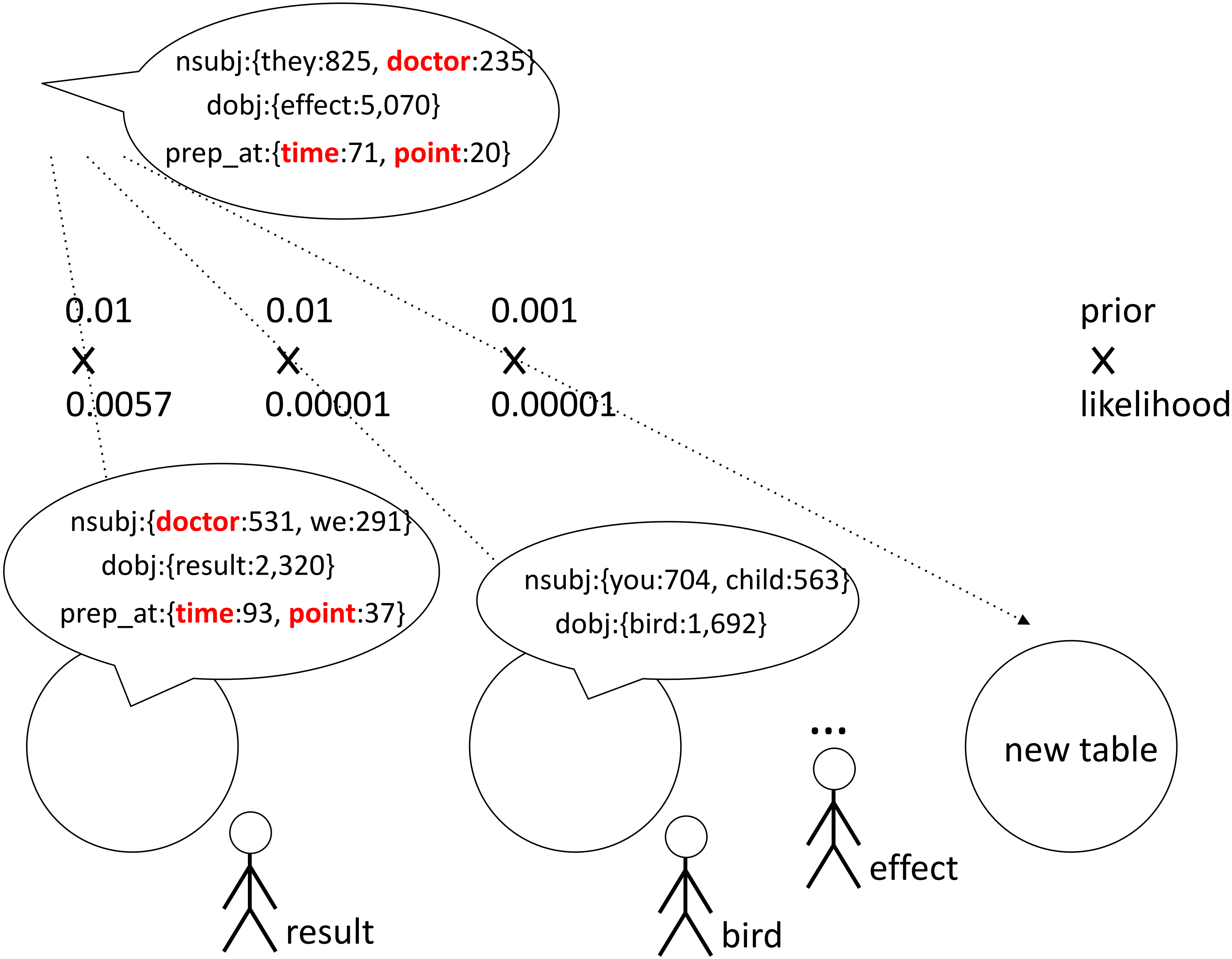
$$P(c_j \mid f_i) \propto \begin{cases} \dfrac{n(c_j)}{N+\alpha} \cdot P(f_i \mid c_j) & j \neq new \\ \dfrac{\alpha}{N+\alpha} \cdot P(f_i \mid c_j) & j = new \end{cases}$$

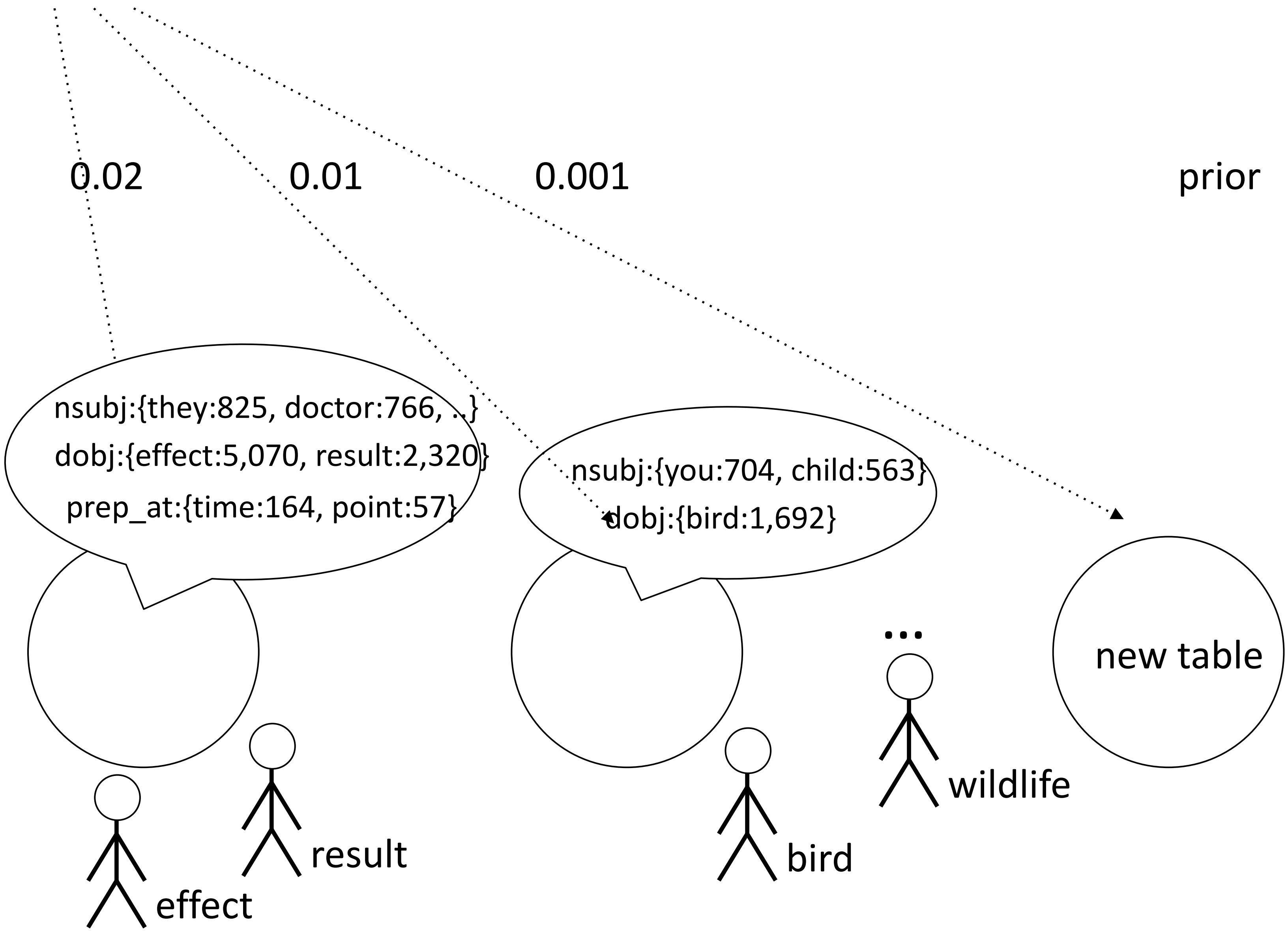
$$P(f_i \mid c_j) = \prod_{w \in W} P(w \mid c_j)^{count(f_i, w)}$$

$$P(w \mid c_j) = \frac{count(c_j, w) + \beta}{\sum_{v \in W} count(c_j, v) + |W| \cdot \beta}$$

f_i : initial frame

c_j : cluster (semantic frame)


nsubj:{they:825, doctor:235}
dobj:{effect:5,070}
prep_at:{time:71, point:20}
0.01
X
0.0057
0.01
X
0.00001
0.001
X
0.00001
prior
X
likelihood
nsubj:{doctor:531, we:291}
dobj:{result:2,320}
prep_at:{time:93, point:37}
nsubj:{you:704, child:563}
dobj:{bird:1,692}
...
new table
result
bird
effect



0.02
0.01
0.001
prior
nsubj:{they:825, doctor:766, ..}
dobj:{effect:5,070, result:2,320}
prep_at:{time:164, point:57}
nsubj:{you:704, child:563}
dobj:{bird:1,692}
...
new table
effect
result
bird
wildlife

nsubj:{they, doctor, ..} observe dobj:{effect}

nsubj.they:825
nsubj.doctor:235

dobj.effect:5,070

prep_at:{time, point, ..}

prep_at.time:71
prep_at.point:20

nsubj:{doctor, we, ..} observe dobj:{result}

nsubj.doctor:531
nsubj.we:291

dobj.result:2,320

prep_at:{time, point, ..}

prep_at.time:93
prep_at.point:37

nsubj:{you, child, ..} observe dobj:{bird}

nsubj.you:704
nsubj.child:563

dobj.bird:1,692

nsubj:{they, doctor, ..} <u>observe</u> dobj:{effect, result}

nsubj.they:825
nsubj.doctor:766
nsubj.we:291

dobj.effect:5,070
dobj.result:2,320

prep_at:{time, point, ..}

prep_at.time:164
prep_at.point:57

nsubj:{you, child, ..} <u>observe</u> dobj:{bird}

nsubj.you:704
nsubj.child:563

dobj.bird:1,692

orchid.kuee.kyoto-u.ac.jp/~kawahara/cf-search/english.crp.gigaword/

Case frame search

Verb: [ ] search reset
Number of displayed instances: 10

"observe" has 16 frames.

ID: observe:1

nsubj	<name>:7296, he:1656, she:449, we:253, they:234, i:187, it:179, report:121, official:114, analyst:110, president:73, ...
ccomp	<s>:14587
prep_in	<name>:100, interview:69, book:32, report:23, speech:19, article:16, 1960s:13, editorial:12, statement:11, commentary:10, scenario:10, ...

ID: observe:2

nsubj	<name>:1227, people:182, they:162, he:156, country:135, muslim:131, nation:125, we:99, state:77, it:76, most:69, ...
dobj	<name>:1221, day:932, anniversary:865, holiday:453, month:227, time:188, birthday:144, ceremony:129, ban:128, festival:102, fast:95, ...
prep_on	<name>:290, day:15, subway:6, spot:6, treatment:6, complex:4, display:4, sale:3, control:3, protection:2, test:2, ...
prep_in	<name>:132, way:12, city:10, honour:9, memory:9, 1675:7, state:6, protest:5, town:4, day:4, month:4, ...

ID: observe:3

nsubj	<name>:1035, group:402, rebel:243, it:240, they:235, side:228, government:196, force:145, troop:114, faction:110, militants:79, ...
dobj	cease-fire:2789, truce:1461, period:355, agreement:283, cessation:48, <name>:47, deal:43, gorilla:40, which:36, suspension:32, border:27, ...
prep_since	<name>:372, 1997:71, 1994:8, end:6, beginning:3, start:3, war:2, eve:2
prep_in	<name>:85, support:38, campaign:12, region:11, territory:9, fight:8, theory:8, war:7, line:6, area:5, republic:5, ...

Acquired from the English Gigaword Corpus (4G words)

Case frame examples for *tsumu* (積む)

	CS	instances (translated into English)
tsumu (1) (accumulate experience)	*ga*	player:21, all:20, person:142, …
	wo	experience:100127, achievement:10350, …
	de	site:240, area:209, …
tsumu (2) (pursue/ devote)	*ga*	person:27, player:13, all:12, …
	wo	exercise:15579, study:13222, …
	de	basis:694, under:384, university:99, …
tsumu (3) (load)	*ga*	man:33, person:20, child:11, …
	wo	baggage:11294, luggage:2989, …
	ni	car:920, truck:160, bike:114, …
…		

ga: nominative, *wo*: accusative, *ni*: dative, *de*: instrument

Table of Contents

- Knowledge for NLP
- Case frame acquisition
- Paraphrase acquisition
- Relation extraction
- Entailment acquisition

Paraphrase in NLP (1/2)

- NLP is difficult because human languages
  - are ambiguous (syntactically, semantically, ...)
  - allow us to express the same information in many greatly different ways, i.e., paraphrase!
- Automatic paraphrase identification/generation
  - Shakespeare is the author of Hamlet.
    - Shakespeare wrote Hamlet. ✓
    - Hamlet is one of Shakespeare's works. ✓
    - Shakespeare was a poet of England. ✗

Paraphrase in NLP (2/2)

- Question answering
  - Questions and their answers in texts are often written in different ways.

  Q: *Who suffers bone fracture?*

  A: *Ken does.*

  ····

  Ken has a broken bone due to...

  ····

  Paraphrase *Ken suffers bone fracture*

- Summarization
  - Redundancy in a text must be identified and removed to summarize it.

  Ken found a solution to a problem. ····

  ······

  ···· *was the problem that Ken solved.* ···

  Paraphrase

Automatic Paraphrase Knowledge Extraction from Texts (1/3)

- Distributional similarity [Lin and Pantel 2001]
  - X is the author of Y
  - X wrote Y
    - X ... Shakespeare, Haruki Murakami, ...
    - Y ... Hamlet, 1Q84, ...
  - X was purchased by Y
  - Y bought X
    - X ... AOL, M&FC, ...
    - Y ... Time Warner Inc., *Nihon Seimitsu Co., Ltd.*, ...
- Difficult to reliably extract paraphrases whose component phrases are infrequent in texts.

Automatic Paraphrase Knowledge Extraction from Texts (2/3)

- Multiple translations of the same text [Barzilay and McKeown 2001]
  - e.g., English translations of the French novel, *Madame Bovary*:
    - Emma burst into tears and he tried to comfort her, saying things to make her smile.
    - Emma cried, and he tried to console her, adorning his words with puns.
- Machine-readable, freely available multiple translations are rare.

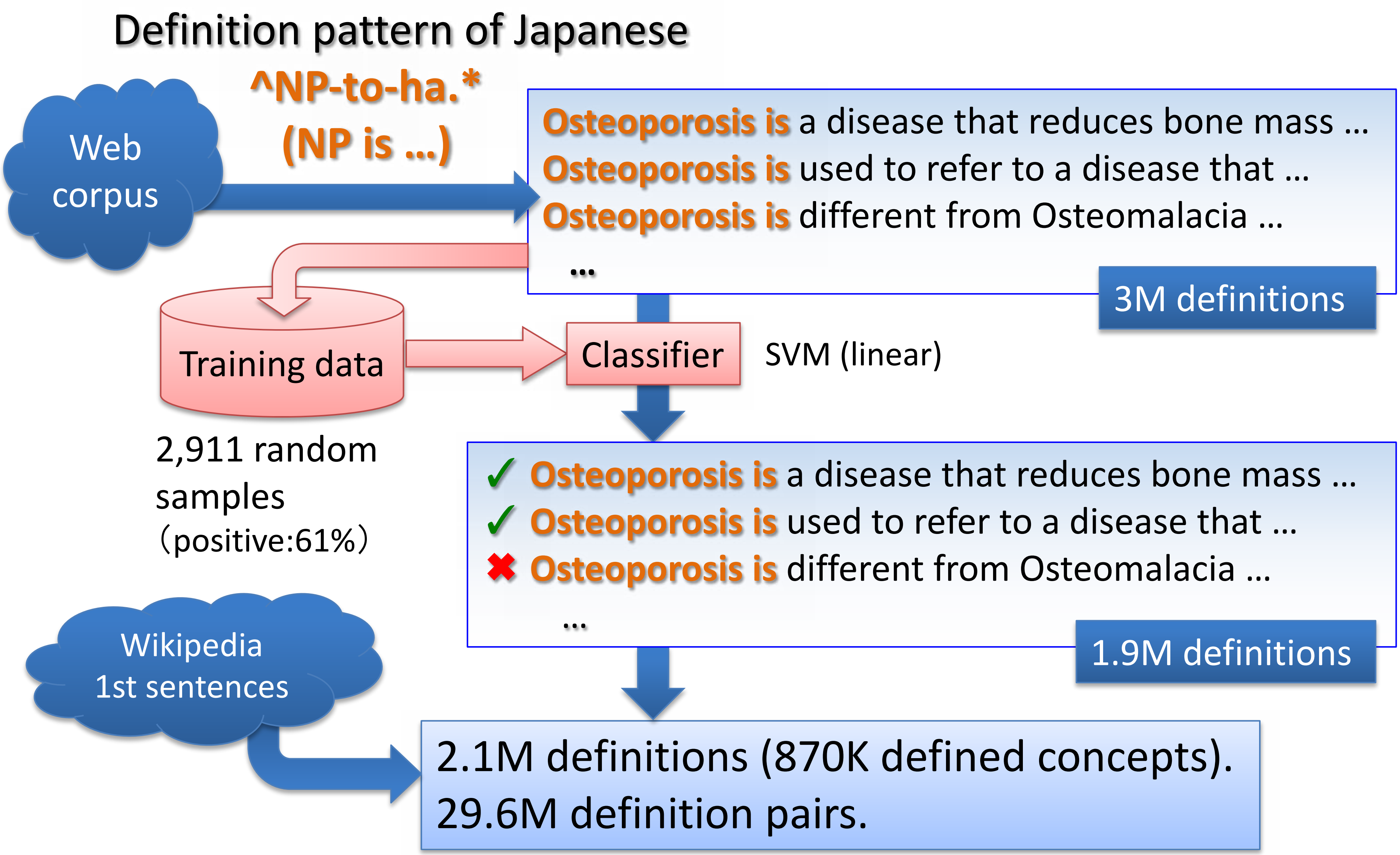
Automatic Paraphrase Knowledge Extraction from Texts (3/3)

- Parallel news sources [Shinyama+ 2003]
  - *New York Times*: Bush, in New York, Affirms $20 Billion Aid Pledge.
  - *Washington Post*: Bush Reassures New York of $20 Billion.
    - PERSON, in LOCATION, affirms MONEY Aid Pledge
    - PERSON reassures LOCATION of MONEY
- Requires a lot of computational cost to acquire them on a large scale.

Paraphrase Knowledge Extraction from Definition Sentences [Hashimoto+ 2011]

- Sentences defining the same thing
  - Osteoporosis is a disease that decreases the quantity of bones and makes bones fragile.
  - Osteoporosis is a disease that reduces bone mass and increases the risk of bone fracture with age.
- Can extract infrequent paraphrases if they appear on definition sentences.
- There are hundreds of millions of definition sentences on the Web.
- Requires no heavy process to acquire definition sentences on the Web.
- About 300K paraphrases with a precision rate of about 94% from a Web corpus of 600M pages.

Definition Sentence Acquisition from the Web (1/2)



Definition Sentence Acquisition from the Web (2/2)

- Features: Bag-of-words and N-grams around the head of sentence and/or right after the definition pattern.
  - ✓ **Osteoporosis is** a disease that reduces bone …
  - ✓ **Osteoporosis is** used to refer to a disease that …
  - ✖ **Osteoporosis is** different from Osteomalacia …
  - Represented by: surface form, base form, POS
- Accuracy: **89.4**, **F1: 91.4**

Examples of Defined Concepts Acquired

Covering a variety of concepts

- FX (Foreign Exchange)
- Metabolic syndrome
- Aegis (Battle ship)
- POCKET MONSTERS (Animated cartoon)
- Caipirinha (Cocktail)
- LOHAS
- Phishing
- Nirvana (Buddhism, Rock ’n’ roll band)
- The Coen Brothers (Film producer)

A variety of paraphrases can be extracted

Paraphrase Extraction from definition pairs (1/3)

- **Osteoporosis** is a disease that **reduces bone mass** and **increases the risk of bone fracture**.
- **Osteoporosis** is a disease that **decreases the quantity of bone** and **makes bones fragile**.

reduces bone mass ⇔ decreases the quantity of bone
reduces bone mass ⇔ makes bones fragile
increases the risk of bone fracture ⇔ decreases the quantity of bone
increases the risk of bone fracture ⇔ makes bones fragile

2,964 pairs

Training data

positive:37%

Classifier

SVM (polynomial 2)

✓ **reduces bone mass ⇔ decreases the quantity of bone**
✖ **reduces bone mass ⇔ makes bones fragile**
✖ **increases the risk of bone fracture ⇔ decreases the quantity of bone**
✓ **increases the risk of bone fracture ⇔ makes bones fragile**

Ranked by the distance from the hyperplane

Paraphrase Extraction from definition pairs (2/3)

Candidate Phrases

- **Osteoporosis** is a disease that **reduces bone mass** and **increases the risk of bone fracture**.
- **Osteoporosis** is a disease that **decreases the quantity of bone** and **makes bones fragile**.

1. Each definition sentence is dependency-parsed.
2. Dependency tree fragments that meet the following conditions become candidate phrases.
   - Consisting of at most 30 words that are consecutive.
   - Containing at least one dependency relation.
   - Headed by verbs, adjectives, or nominal predicates.
   - Containing no demonstratives.

Paraphrase Extraction from definition pairs (3/3)

- Feature set 1: Similarity between candidate phrases
- Feature set 2: Similarity between their contexts

Context	Candidate phrase	Context
Osteoporosis is a disease that	reduces bone mass	and makes bones fragile.
Osteoporosis is a disease that	decreases the quantity of bone	and makes bones easy to fracture.

Feature set 1:

- Word overlap
- Semantically similar words
- Identity of head word
  - Inflected form, POS, ⋯
- ・・・

Feature set 2:

- N-gram overlap
  - Base form, pronunciation
- Dependency tree fragment overlap
  - POS, base form, pronunciation
- ・・・

Examples of Extracted Paraphrases

Note: The target language is Japanese. Examples are translated from Japanese results.

- cause the oxidation of cells ⇔ cause cellular aging
- correct eyesight ⇔ perform eyesight correction
- access Web sites ⇔ visit WWW sites
- trade foreign currencies ⇔
  exchange one currency for another
- mount two processor cores on one CPU ⇔
  build two processor cores into one package

Examples of Extracted Incorrect Paraphrases

- send to a Web browser ⇔ send to a PC
- intend to balance ⇔ intend to refresh
- unable to digest with digestive enzymes ⇔ hard to digest with digestive enzymes

A More Ambitious Hypothesis

- Sentences fulfilling the same function for the same topic are paraphrases of each other?
  - Definition of the same concept ✓
  - Usage of the same Linux command
  - Recipe for the same cuisine
  - Description of related work on the same research issue
  - …

Table of Contents

- Knowledge for NLP
- Case frame acquisition
- Paraphrase acquisition
- Relation extraction
- Entailment acquisition

Relation Extraction

- X has a relation to Y: (X <rel> Y)

- For some typical relations like <is a>, we can make patterns:
  - "X such as Y and Z" → (Y <is a> X), (Z <is a> X)

- However, hand-crafting rules is not practical, considering long-tail relations and long-tail patterns

Espresso [Pantel and Pennacchiotti 2006]

1. Give some seed instances
2. Learn patterns that co-occur with instances
3. Apply patterns to get new instances

$$r_p(p) = \frac{\sum_{i \in I} \{ \frac{\mathrm{pmi}(i, p)}{\max_{pmi}} \times r_i(i) \}}{|I|} \quad r_i(i) = \frac{\sum_{p \in P} \{ \frac{\mathrm{pmi}(i, p)}{\max_{pmi}} \times r_p(p) \}}{|P|}$$

Espresso

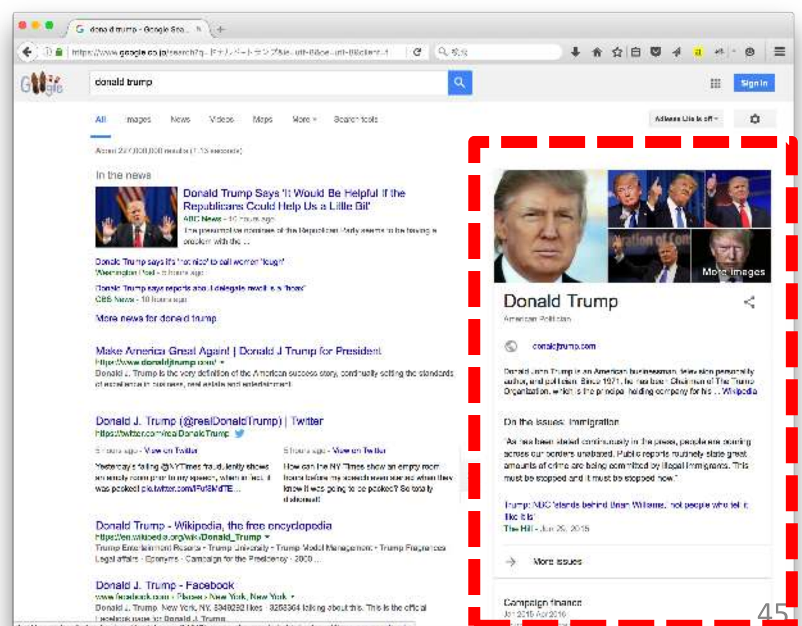
Table 1. Sample seeds used for each semantic relation and sample outputs from *Espresso*. The number in the parentheses for each relation denotes the total number of seeds used as input for the system.

	Is-a (12)	*Part-Of (12)*	*Succession (12)*	*Reaction (13)*	*Production (14)*
Seeds	wheat :: crop George Wendt :: star nitrogen :: element diborane :: substance	leader :: panel city :: region ion :: matter oxygen :: water	Khrushchev :: Stalin Carla Hills :: Yeutter Bush :: Reagan Julio Barbosa :: Mendes	magnesium :: oxygen hydrazine :: water aluminum metal :: oxygen lithium metal :: fluorine gas	bright flame :: flares hydrogen :: metal hydrides ammonia :: nitric oxide copper :: brown gas
Es-presso	Picasso :: artist tax :: charge protein :: biopolymer HCl :: strong acid	trees :: land material :: FBI report oxygen :: air atom :: molecule	Ford :: Nixon Setrakian :: John Griesemer Camero Cardiel :: Camacho Susan Weiss :: editor	hydrogen :: oxygen Ni :: HCl carbon dioxide :: methane boron :: fluorine	electron :: ions glycerin :: nitroglycerin kidneys :: kidney stones ions :: charge

Knowledge Bases

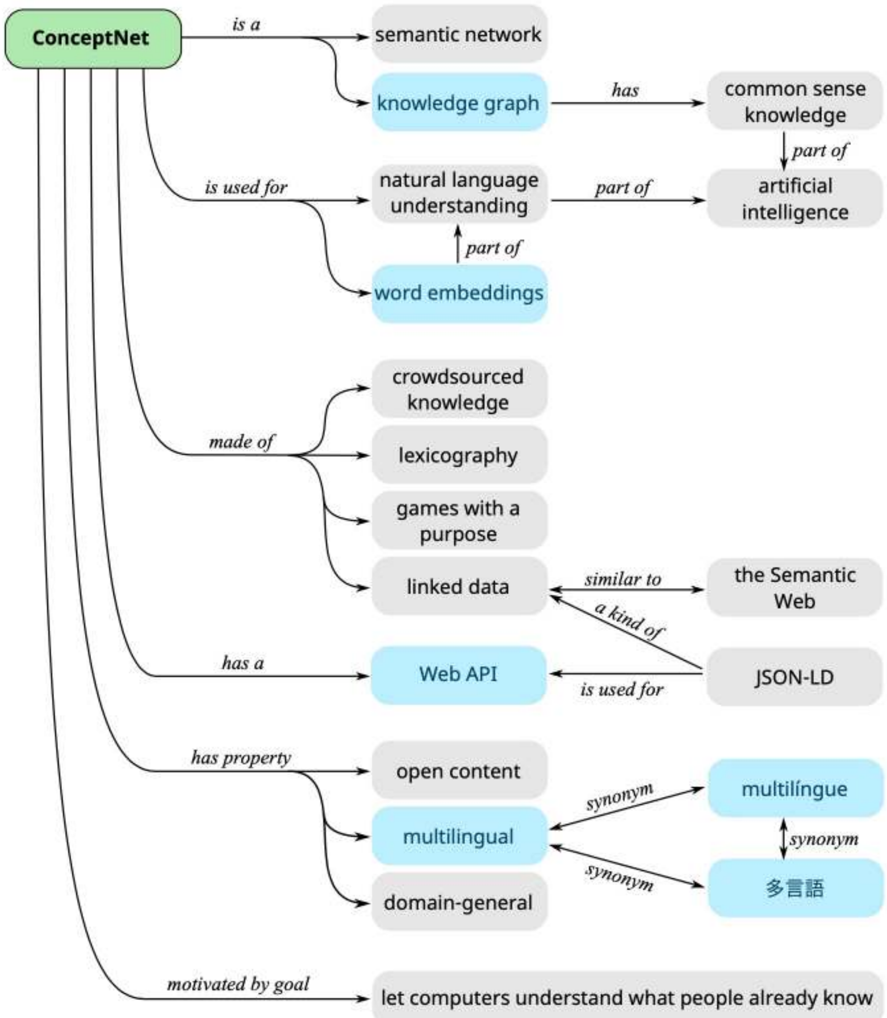
- DBPedia https://www.dbpedia.org/
  – Automatically extracted relations mainly from Wikipedia Infobox
  – Over 6M entities and 9.5 billion triples (as of 2016)

- Knowledge Graph
  – Over 5 billion entities and over 500 billion relations (as of 2020)
  – Gathered from various sources



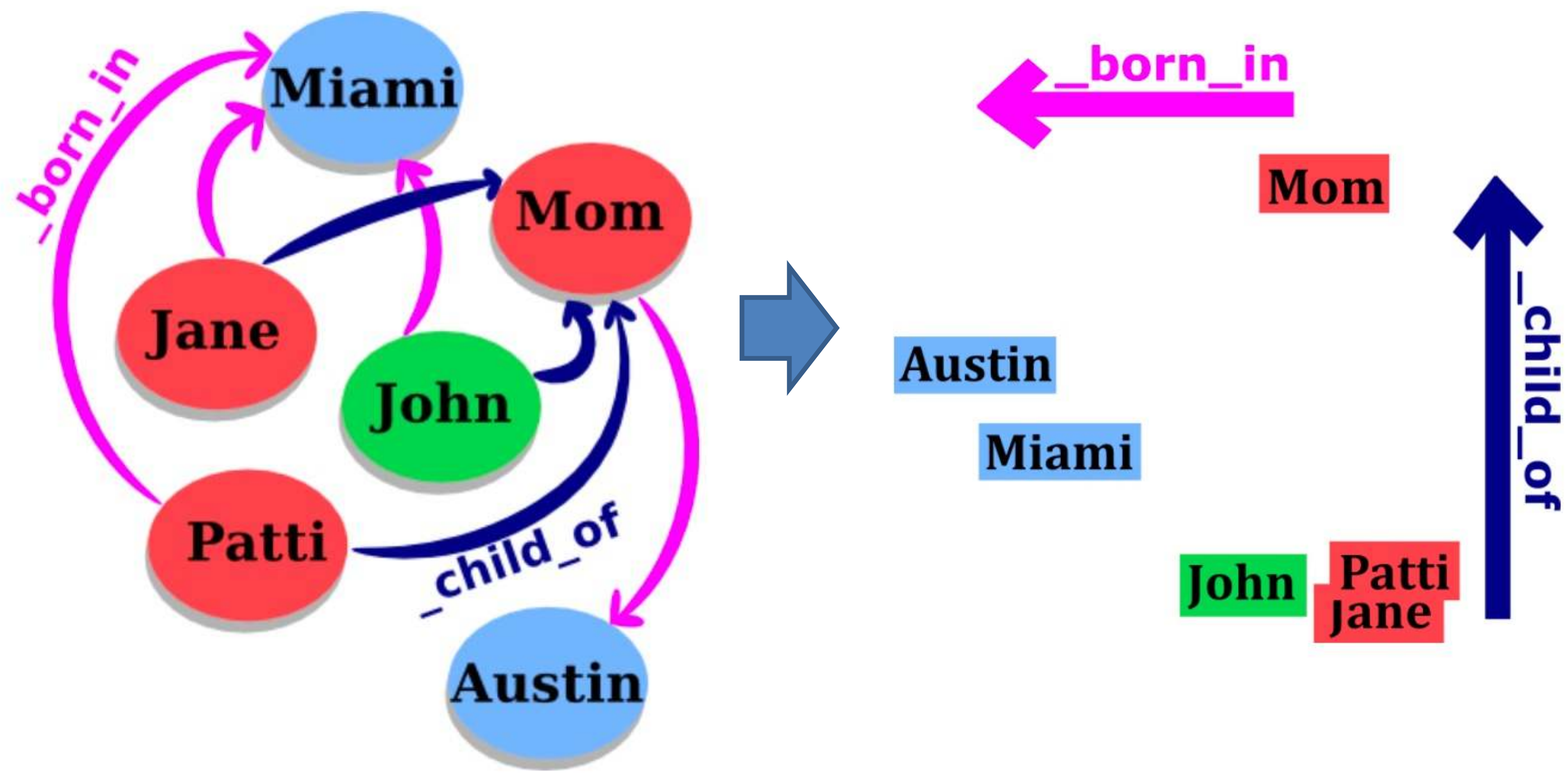
Knowledge Bases

- Wikidata https://www.wikidata.org/
  - Collaboratively edited knowledge graph
  - 94M items and 12 billion statements
- ConceptNet [Speer+ 2017]
  - Concepts are expressed in natural language
  - 8M nodes and 21M edges



https://conceptnet.io/

TransE [Bordes+ 2013]



http://cilvr.cs.nyu.edu/lib/exe/fetch.php?media=deeplearning:2015:dl-nyu-bordes.pdf

TransE [Bordes+ 2013]

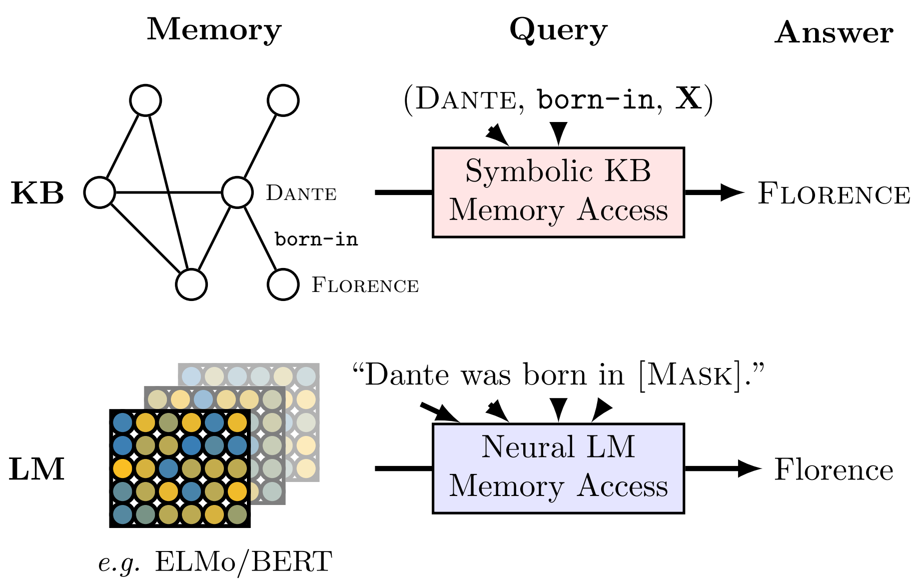
- Intuition: we want $h + r \approx t$
- Training: minimizing $\mathcal{L}$

$$\mathcal{L} = \sum_{pos} \sum_{neg \in S'} [\gamma + d(h + r, t) - d(h' + r, t')]_+$$

where $[x]_+$ is the positive part of x, $\gamma > 0$ is a margin, and
$S' = \{(h', r, t) | h' \in E\} \cup \{(h, r, t') | t' \in E\}$.

Language Models as Knowledge Bases

[Petroni+ 2019]



Corpus	Relation	Statistics #Facts	#Rel	Baselines Freq	DrQA	KB RE_n	RE_o	LM Fs	Txl	Eb	E5B	Bb	Bl
Google-RE	`birth-place`	2937	1	4.6	-	3.5	13.8	4.4	2.7	5.5	7.5	14.9	**16.1**
	`birth-date`	1825	1	1.9	-	0.0	**1.9**	0.3	1.1	0.1	0.1	1.5	1.4
	`death-place`	765	1	6.8	-	0.1	7.2	3.0	0.9	0.3	1.3	13.1	**14.0**
	Total	5527	3	4.4	-	1.2	7.6	2.6	1.6	2.0	3.0	9.8	**10.5**
T-REx	1-1	937	2	1.78	-	0.6	10.0	17.0	36.5	10.1	13.1	68.0	**74.5**
	N-1	20006	23	23.85	-	5.4	**33.8**	6.1	18.0	3.6	6.5	32.4	34.2
	N-*M*	13096	16	21.95	-	7.7	**36.7**	12.0	16.5	5.7	7.4	24.7	24.3
	Total	34039	41	22.03	-	6.1	**33.8**	8.9	18.3	4.7	7.1	31.1	32.3
ConceptNet	Total	11458	16	4.8	-	-	-	3.6	5.7	6.1	6.2	15.6	**19.2**
SQuAD	Total	305	-	-	**37.5**	-	-	3.6	3.9	1.6	4.3	14.1	17.4

Language Models as Knowledge Bases

[Petroni+ 2019]

	Relation	Query	Answer	Generation
T-Rex	P19	Francesco Bartolomeo Conti was born in ____.	Florence	Rome [-1.8] , **Florence [-1.8]** , Naples [-1.9] , Milan [-2.4] , Bologna [-2.5]
	P20	Adolphe Adam died in ____.	Paris	**Paris [-0.5]** , London [-3.5] , Vienna [-3.6] , Berlin [-3.8] , Brussels [-4.0]
	P279	English bulldog is a subclass of ____.	dog	dogs [-0.3] , breeds [-2.2] , **dog [-2.4]** , cattle [-4.3] , sheep [-4.5]
	P37	The official language of Mauritius is ____.	English	**English [-0.6]** , French [-0.9] , Arabic [-6.2] , Tamil [-6.7] , Malayalam [-7.0]
	P413	Patrick Oboya plays in ____ position.	midfielder	centre [-2.0] , center [-2.2] , **midfielder [-2.4]** , forward [-2.4] , midfield [-2.7]
	P138	Hamburg Airport is named after ____.	Hamburg	Hess [-7.0] , Hermann [-7.1] , Schmidt [-7.1] , **Hamburg [-7.5]** , Ludwig [-7.5]
	P364	The original language of Mon oncle Benjamin is ____.	French	**French [-0.2]** , Breton [-3.3] , English [-3.8] , Dutch [-4.2] , German [-4.9]
	P54	Dani Alves plays with ____ .	Barcelona	Santos [-2.4] , Porto [-2.5] , Sporting [-3.1] , Brazil [-3.3] , Portugal [-3.7]
	P106	Paul Toungui is a ____ by profession .	politician	lawyer [-1.1] , journalist [-2.4] , teacher [-2.7] , doctor [-3.0] , physician [-3.7]
	P527	Sodium sulfide consists of ____.	sodium	water [-1.2] , sulfur [-1.7] , **sodium [-2.5]** , zinc [-2.8] , salt [-2.9]
	P102	Gordon Scholes is a member of the ____ political party.	Labor	Labour [-1.3] , Conservative [-1.6] , Green [-2.4] , Liberal [-2.9] , **Labor [-2.9]**
	P530	Kenya maintains diplomatic relations with ____.	Uganda	India [-3.0] , **Uganda [-3.2]** , Tanzania [-3.5] , China [-3.6] , Pakistan [-3.6]
	P176	iPod Touch is produced by ____.	Apple	**Apple [-1.6]** , Nokia [-1.7] , Sony [-2.0] , Samsung [-2.6] , Intel [-3.1]
	P30	Bailey Peninsula is located in ____.	Antarctica	**Antarctica [-1.4]** , Bermuda [-2.2] , Newfoundland [-2.5] , Alaska [-2.7] , Canada [-3.1]
	P178	JDK is developed by ____.	Oracle	IBM [-2.0] , Intel [-2.3] , Microsoft [-2.5] , HP [-3.4] , Nokia [-3.5]
	P1412	Carl III used to communicate in ____.	Swedish	German [-1.6] , Latin [-1.9] , French [-2.4] , English [-3.0] , Spanish [-3.0]
	P17	Sunshine Coast, British Columbia is located in ____.	Canada	**Canada [-1.2]** , Alberta [-2.8] , Yukon [-2.9] , Labrador [-3.4] , Victoria [-3.4]
	P39	Pope Clement VII has the position of ____ .	pope	cardinal [-2.4] , Pope [-2.5] , **pope [-2.6]** , President [-3.1] , Chancellor [-3.2]
	P264	Joe Cocker is represented by music label ____.	Capitol	EMI [-2.6] , BMG [-2.6] , Universal [-2.8] , **Capitol [-3.2]** , Columbia [-3.3]
	P276	London Jazz Festival is located in ____.	London	**London [-0.3]** , Greenwich [-3.2] , Chelsea [-4.0] , Camden [-4.6] , Stratford [-4.8]
	P127	Border TV is owned by ____.	ITV	Sky [-3.1] , **ITV [-3.3]** , Global [-3.4] , Frontier [-4.1] , Disney [-4.3]
	P103	The native language of Mammootty is ____.	Malayalam	**Malayalam [-0.2]** , Tamil [-2.1] , Telugu [-4.8] , English [-5.2] , Hindi [-5.6]
	P495	The Sharon Cuneta Show was created in ____.	Philippines	Manila [-3.2] , **Philippines [-3.6]** , February [-3.7] , December [-3.8] , Argentina [-4.0]
ConceptNet	AtLocation	You are likely to find a overflow in a ____.	drain	sewer [-3.1] , canal [-3.2] , toilet [-3.3] , stream [-3.6] , **drain [-3.6]**
	CapableOf	Ravens can ____.	fly	**fly [-1.5]** , fight [-1.8] , kill [-2.2] , die [-3.2] , hunt [-3.4]
	CausesDesire	Joke would make you want to ____.	laugh	cry [-1.7] , die [-1.7] , **laugh [-2.0]** , vomit [-2.6] , scream [-2.6]
	Causes	Sometimes virus causes ____.	infection	disease [-1.2] , cancer [-2.0] , **infection [-2.6]** , plague [-3.3] , fever [-3.4]
	HasA	Birds have ____.	feathers	wings [-1.8] , nests [-3.1] , **feathers [-3.2]** , died [-3.7] , eggs [-3.9]
	HasPrerequisite	Typing requires ____.	speed	patience [-3.5] , precision [-3.6] , registration [-3.8] , accuracy [-4.0] , **speed [-4.1]**
	HasProperty	Time is ____.	finite	short [-1.7] , passing [-1.8] , precious [-2.9] , irrelevant [-3.2] , gone [-4.0]
	MotivatedByGoal	You would celebrate because you are ____.	alive	happy [-2.4] , human [-3.3] , **alive [-3.3]** , young [-3.6] , free [-3.9]
	ReceivesAction	Skills can be ____.	taught	acquired [-2.5] , useful [-2.5] , learned [-2.8] , combined [-3.9] , varied [-3.9]
	UsedFor	A pond is for ____.	fish	swimming [-1.3] , fishing [-1.4] , bathing [-2.0] , **fish [-2.8]** , recreation [-3.1]

Table of Contents

- Knowledge for NLP
- Case frame acquisition
- Paraphrase acquisition
- Relation extraction
- Entailment acquisition

Textual Entailment

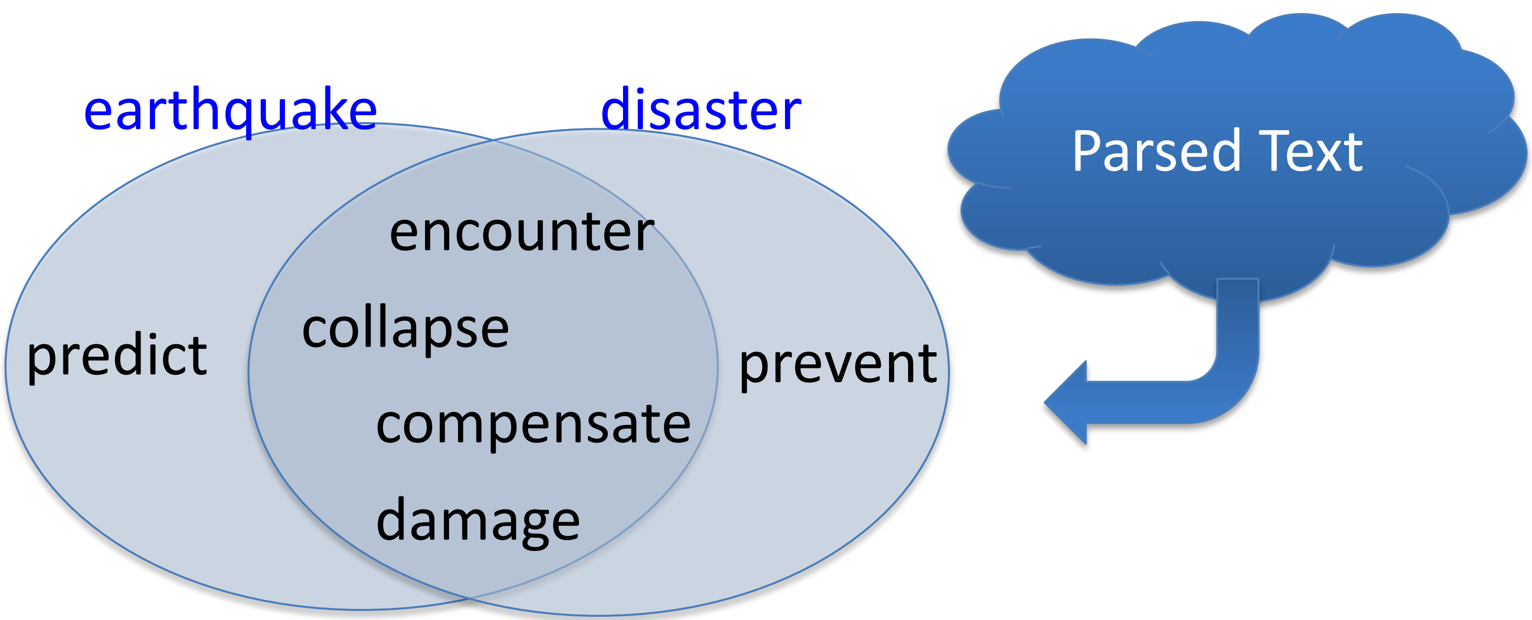
- T entails H1 and H2, but does not entail H3.
  - T: Taro, a student of Waseda University, attended the Natural Language Processing class and got tired of it.
  - H1: Taro entered Waseda University.
  - H2: Taro studies Natural Language Processing.
  - H3: Taro likes the Natural Language Processing class.
- **T entails H if H is true whenever T is true.**

Verb Entailment

- Left side verbs entail right side verbs.
  - microwave → warm
  - commit apoptosis → die
  - gulp → drink
  - snore → sleep
  - divorce → marry
  - fall off a horse → ride a horse
- Left side verbs do NOT entail right side verbs.
  - commute → take a transfer
  - scream → be surprised
  - get sick → be cured
- ***verb*1 entails *verb*2 if *verb*1 cannot be done unless *verb*2 is, or has been, done.**

Distributional Similarity

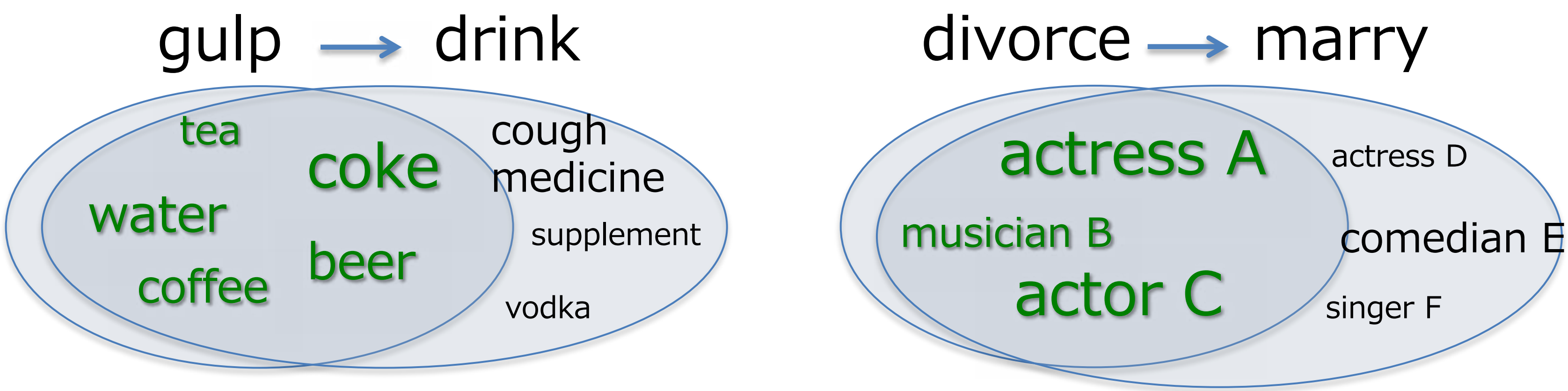
- Assumption: Two words that appear in similar contexts tend to be semantically similar.
  1. Parse a large corpus.
  2. For each target word, obtain its contexts from the parsed corpus.
  3. For each word pair, compare their contexts and measure the size of overlap between their contexts.



Directional Distributional Similarity

[Hashimoto+ 2011]

- Assumption: If the context of verb1 is subsumed by that of verb2, verb1 entails verb2.

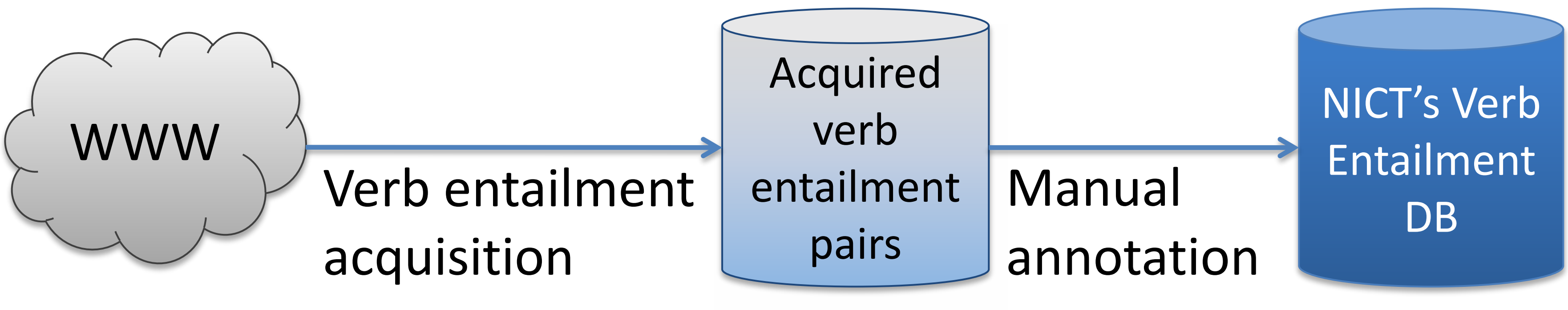


Weighting Context Words

- Context words are not equally important.
  - Some context words of a target word show the target's characteristics more strongly.
  - "nurse"
    - more important: "patient", "child"
    - less important: "Mr. Smith", "him"
- Point-wise Mutual Information is effective.

$$\boxed{\log \frac{p(nurse, patient)}{p(nurse) \cdot p(patient)}}$$

NICT’s Verb Entailment Database



- Positive examples：50,079 pairs
  - microwave → warm
  - commit apoptosis → die
  - gulp → drink
- Negative examples：38,787 pairs
  - commute → take a transfer
  - scream → be surprised
  - get sick → be cured

The Stanford Natural Language Inference (SNLI) Corpus [Bowman+ 2015]

A collection of 570k human-written English sentence pairs manually labeled for balanced classification with the labels: *entailment*, *contradiction*, and *neutral*

Text	Judgments	Hypothesis
A man inspects the uniform of a figure in some East Asian country.	contradiction C C C C C	The man is sleeping
An older and younger man smiling.	neutral N N E N N	Two men are smiling and laughing at the cats playing on the floor.
A black race car starts up in front of a crowd of people.	contradiction C C C C C	A man is driving down a lonely road.
A soccer game with multiple males playing.	entailment E E E E E	Some men are playing a sport.
A smiling costumed woman is holding an umbrella.	neutral N N E C N	A happy woman in a fairy costume holds an umbrella.

https://nlp.stanford.edu/projects/snli/

The Multi-Genre Natural Language Inference (MultiNLI) Corpus

[Williams+ 2018]

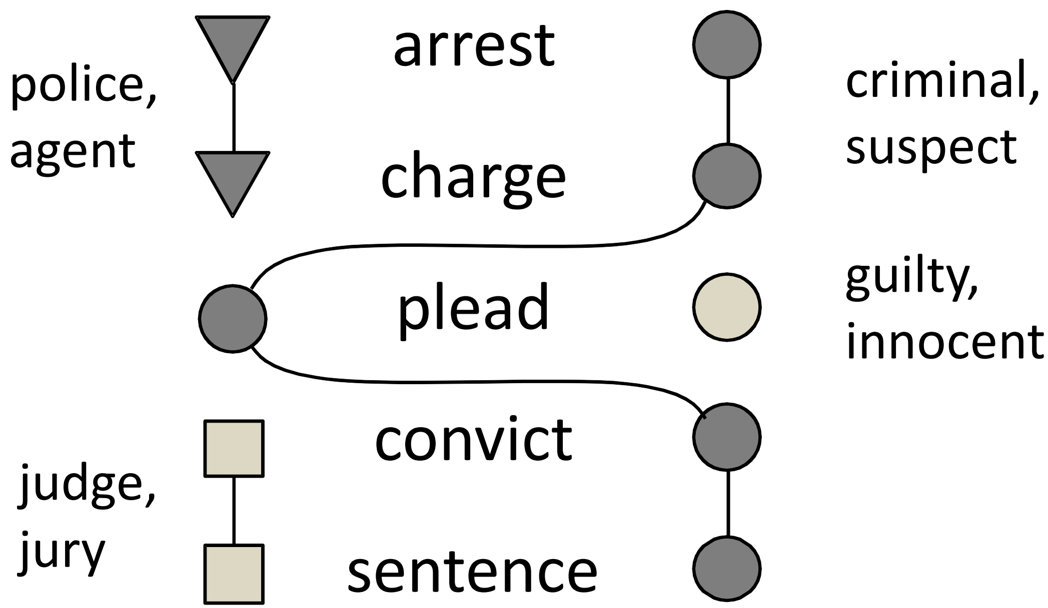
A collection of 433k human-written English sentence pairs on multiple genres manually labeled with *entailment*, *contradiction*, and *neutral*

Met my first girlfriend that way.	FACE-TO-FACE **contradiction** C C N C	I didn't meet my first girlfriend until later.
He turned and saw Jon sleeping in his half-tent.	FICTION **entailment** N E N N	He saw Jon was asleep.
8 million in relief in the form of emergency housing.	GOVERNMENT **neutral** N N N N	The 8 million dollars for emergency housing was still not enough to solve the problem.
Now, as children tend their gardens, they have a new appreciation of their relationship to the land, their cultural heritage, and their community.	LETTERS **neutral** N N N N	All of the children love working in their gardens.
At 8:34, the Boston Center controller received a third transmission from American 11	9/11 **entailment** E E E E	The Boston Center controller got a third transmission from American 11.

Other Topics

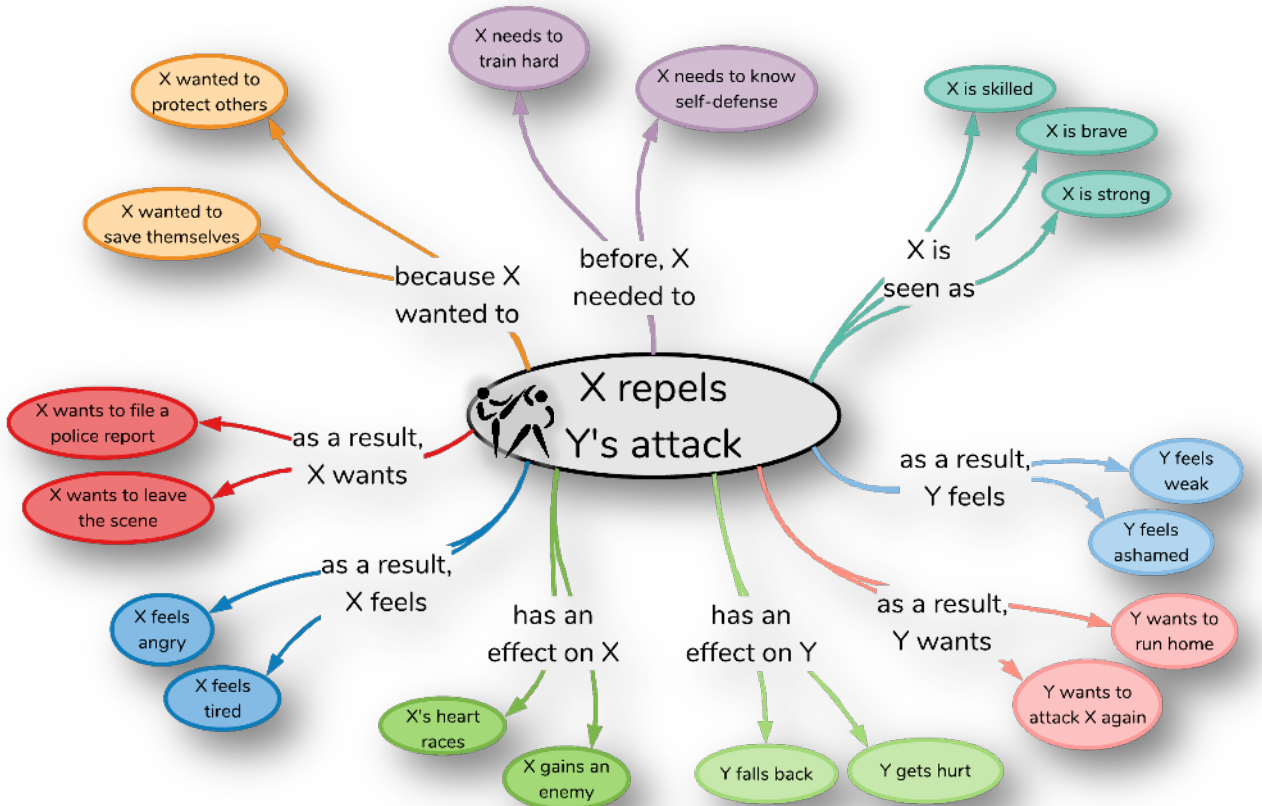
Script

- Unsupervised Learning of Narrative Schemas and their Participants [Chambers+ 2009]



ATOMIC [Sap+ 2019]

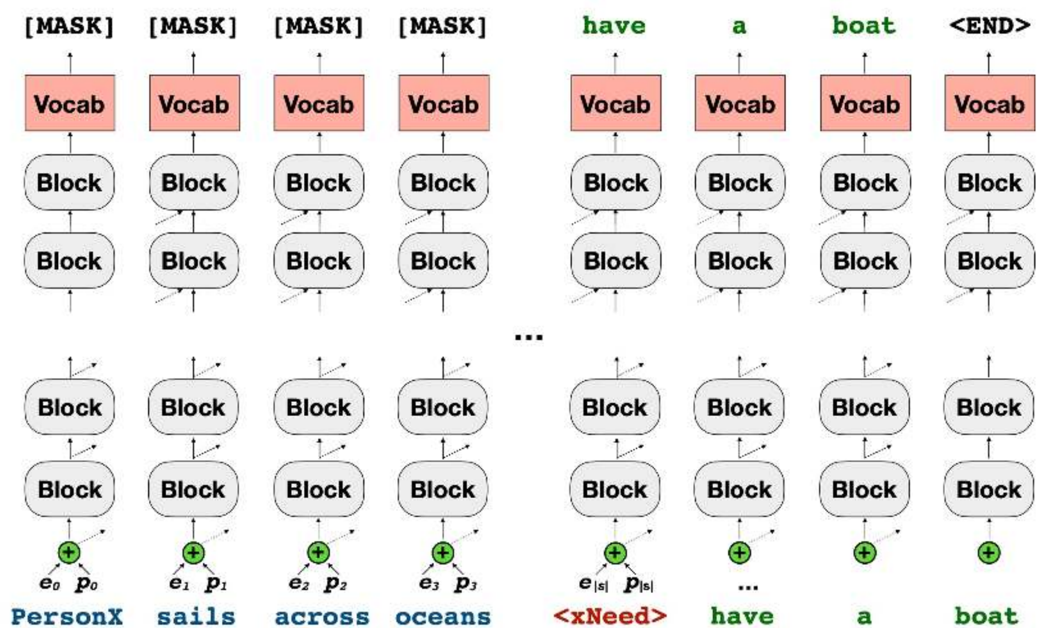
- 880k event-to-event relations obtained by crowdsourcing



COMET [Bosselut+ 2019]

Combined pre-trained language models and knowledge graphs

Up close: Given a **phrase subject** and a **relation**, learn to generate the **phrase object**

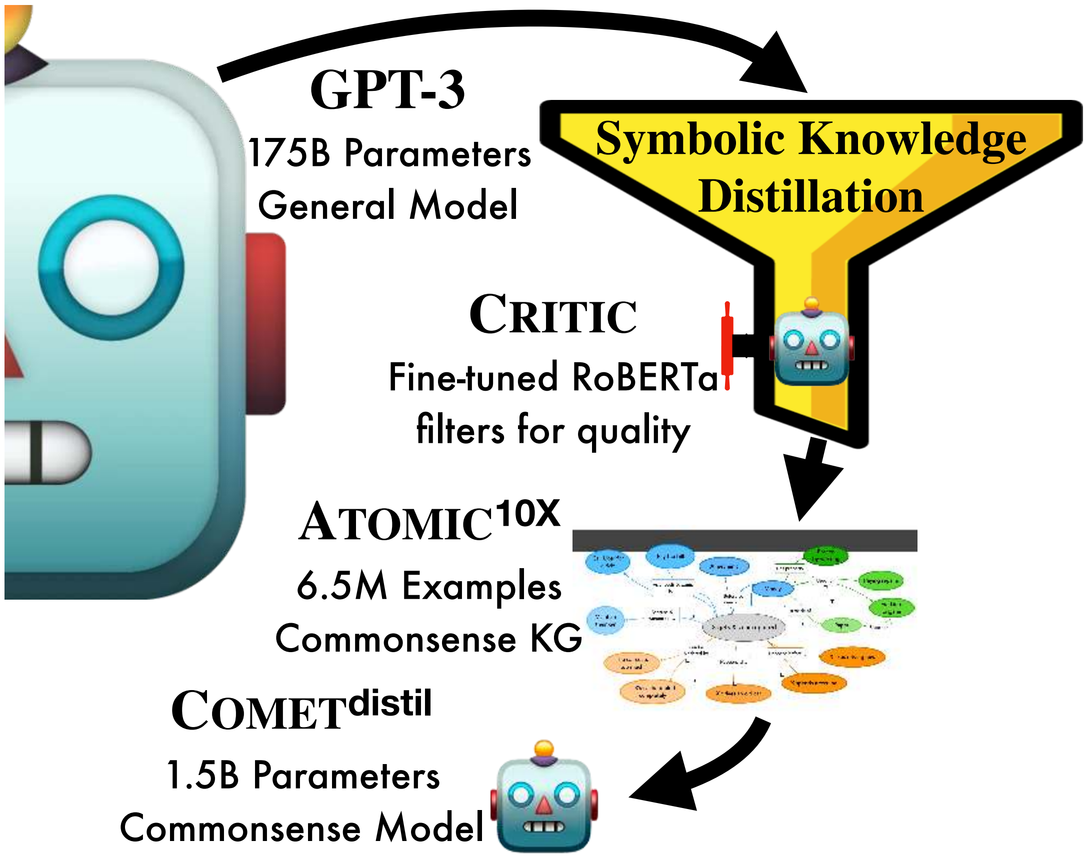


Training with ATOMIC

Seed Concept	Relation	Generated	Plausible
X holds out X's hand to Y	xAttr	helpful	✓
X meets Y eyes	xAttr	intense	✓
X watches Y every ___	xAttr	observant	✓
X eats red meat	xEffect	gets fat	✓
X makes crafts	xEffect	gets dirty	✓
X turns X's phone	xEffect	gets a text	
X pours ___ over Y's head	oEffect	gets hurt	✓
X takes Y's head off	oEffect	bleeds	✓
X pisses on Y's bonfire	oEffect	gets burned	
X spoils somebody rotten	xIntent	to be mean	
X gives Y some pills	xIntent	to help	✓
X provides for Y's needs	xIntent	to be helpful	✓
X explains Y's reasons	xNeed	to know Y	✓
X fulfils X's needs	xNeed	to have a plan	✓
X gives Y everything	xNeed	to buy something	✓
X eats pancakes	xReact	satisfied	✓
X makes ___ at work	xReact	proud	✓
X moves house	xReact	happy	✓
X gives birth to the Y	oReact	happy	✓
X gives Y's friend ___	oReact	grateful	✓
X goes ___ with friends	oReact	happy	✓
X gets all the supplies	xWant	to make a list	✓
X murders Y's wife	xWant	to hide the body	✓
X starts shopping	xWant	to go home	✓
X develops Y theory	oWant	to thank X	✓
X offer Y a position	oWant	to accept the job	✓
X takes ___ out for dinner	oWant	to eat	✓

Symbolic Knowledge Distillation

[West+ 2022]



Symbolic Knowledge Distillation

- Event generation

```
1.   Event:   X overcomes evil with good
2.   Event:   X does not learn from Y
...
10.   Event:   X looks at flowers
11.
```

Example prompt

- Inference generation

```
What needs to be true for this
event to take place?

...

Event <i>:   X goes jogging

Prerequisites:   For this to
happen, X needed to wear running
shoes

...

Event <i>:   X looks at flowers

Prerequisites:   For this to
happen,
```

Example prompt (xNeed)

Winograd Schema Challenge

[Jevesque 2011]

- A dataset for the task of resolving definite pronouns (2,000 problems)
- Require the use of world knowledge and reasoning

The red team defeated the blue team

?

because they made the last penalty kick.

X makes a penalty kick → X defeats Y

Assignment

Select an NLP task or service and report what kind of knowledge is necessary to improve it. Also describe what methods can be used to acquire such knowledge.

Deadline: June 23 (Thu) 23:59

※ You can write it in English or Japanese.

Summary

- Knowledge for NLP
- Case frame acquisition
- Paraphrase acquisition
- Relation extraction
- Entailment acquisition