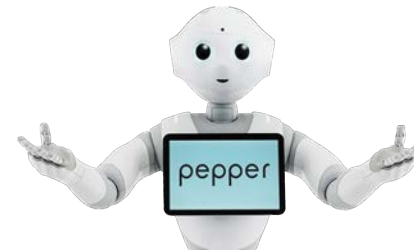東京工業大学
Tokyo Institute of Technology

Annotation Behaviour Mining for Natural Language Processing

2017.3.7@Tor Vergata

Tokunaga, Takenobu
Tokyo Institute of Technology
School of Computing

東京工業大学
Tokyo Institute of Technology

Recently …

2

東京工業大学
Tokyo Institute of Technology

ELIZA (Joseph Weizenbaum 1966)

H: Men are all alike.

E: *In what way?*

H: They're always bugging us about something.

E: *Can you think of a specific example?*

H: Well, my boyfriend made me come here.

E: *Your boyfriend made you come here.*

H: He says I'm depressed much of the time

E: *I am sorry to hear you are depressed.*

…

3

東京工業大学
Tokyo Institute of Technology

SHRDLU (Terry Winograd 1972)

H: How many blocks are not in the box?
S: *Four of them.*
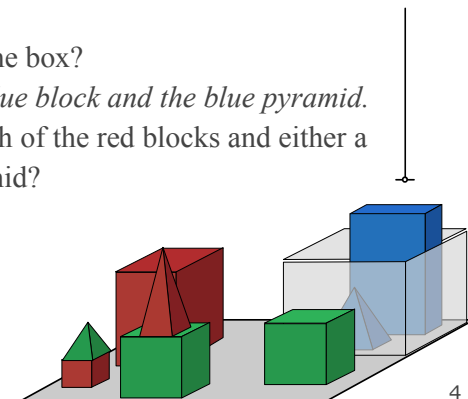
…

H: Is there anything in the box?
S: *Yes, two things: the blue block and the blue pyramid.*
H: Will you stack up both of the red blocks and either a green cube or a pyramid?
S: *OK.*

4

Past (~2000)

1946	ENIAC (MT~)
1957	*Syntactic Structure* (Chomsky 1957)
1962	Association for Machine Translation and Computational Linguistics (AMTCL)
1966	ALPAC report (Pierce 1966)
	ELIZA (Weizenbaum 1966)
1968	Association for Computational Linguistics (ACL)
1972	SHRDLU (Winograd 1972)
1982	ICOT (~ 1992)
1984	EDR (~ 2002)
1986	WordNet (Miller 1995)
1993	Penn Treebank (Marcus 1993)
1999	*Foundations of SNLP* (Manning&Schütze 1999)

Present
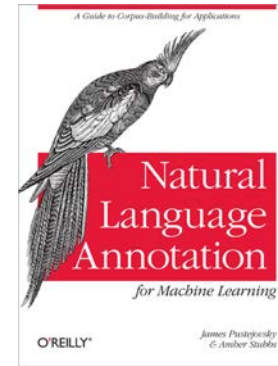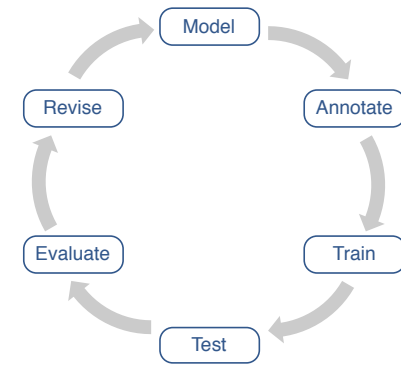
2012   MATTER cycle

Rationalism vs. Empiricism

- Rationalism (1960's~)
  - knowledge construction
    (grammar, lexicon, ontology)
  - symbol calculation on knowledge
- Empiricism (1990's~)
  - corpus annotation → training data
  - machine learning → evaluation

Oppositions

Empiricism	Rationalism
result of production	process of production
autonomous object	contextual dependency
macroscopic	microscopic
big data	small data

Role of corpora in Empiricism

- training data for ML
  - output: annotated information
  - input: features (e.g. surrounding words, POS, distance, ⋯)
- evaluating systems

When it does not work ⋯

- bad annotation? (output)
- bad features? (input)

annotating annotation with its annotation process (meta annotation)

Meta annotation

- on what? → annotation instance
- what? → annotator behaviour
- how? → observation vs. annotator's think aloud

Collecting annotation behaviour

- annotator does
  1. read a text
  2. decide how they annotate on what
  3. annotate it
- collected data
  - observable data (1, 3): eye gaze, tool operation
 - implicit data (1, 2): decision process

Data collection

- observable data
  - eye gaze → eye-tracker
  - tool operation → annotation tool
- implicit data
  - TAP (think-aloud protocol)
  - dialogue analysis ← pair annotation

What collected data is useful for?

- discovery of useful information for NLP
  what information does human use for the NLP task?
- evaluation of annotation quality
  does behaviour imply resultant annotation quality?
- evaluation of annotator ability
  how do expert annotators behave?
- training annotators
  does teaching good behaviour to novice annotator
  improve their annotation skills?

To sum

Current framework
(corpus+ML)
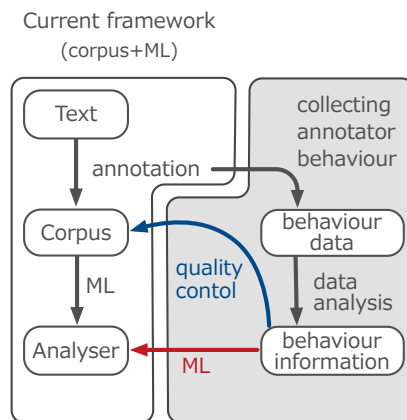
Research items

- Data collection environment
- Annotation tasks
- Annotation pairs
- Integration of observable and implicit data
- Representation of meta annotation

Related work (eye tracking)

- eye tracking during reading and generating texts (Griffin&Bock 2000), (Richardson+ 2007)
- pair programming (Lui&Chan 2006)
- eye tracking during debugging programs (Pietinen+ 2008), (Bednarik&Tukianinen 2008)
- gaze scribing (Rosengrant 2010)
- pair annotation (Demirsahin+ 2012)

Relate work (NLP)

- dialogue system (Prasov+ 2007, Qu&Chai 2007)
- reference resolution (Prasov&Chai 2008)
- named entity annotation (Tomanek+ 2010)
- syntactic analysis (Barrett&Søgaard 2015)
- coreference analysis (Ross+ 2016)

Annotation tasks

- **segmentation task**: identifying text spans
  e.g. POS tagging, NE recognition
- **linking task**: linking between multiple text spans
  e.g. dependency analysis, reference analysis, predicate-argument analysis
- **transformation task**: transform text to text
  e.g. translation, summarisation, paraphrasing

Segmentation task: NE recognition

- identify proper nouns of specific semantic categories in texts, e.g. person name, location name, organisation name …
  → IE, sentiment analysis
- 72 instances that a machine failed to analyse
- participants: 16
- eye tracker: Tobii T-60

Workflow
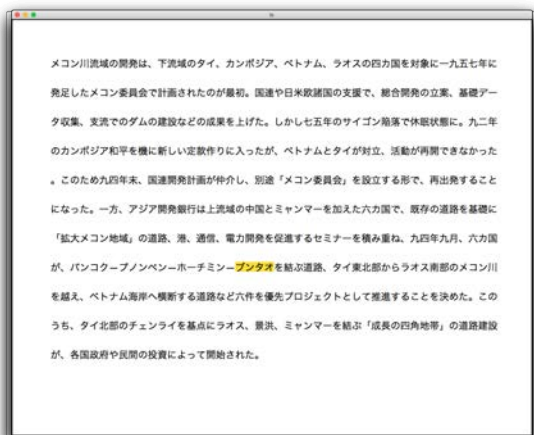
Results

• accuracy of each annotator

anno. ID	01	02	04	08	10	11	12	13	14	15	ave.
#correct	59	54	63	57	64	66	64	47	59	66	59.9
accuracy	.82	.75	.88	.79	.89	.92	.89	.65	.82	.92	.83

• NE difficulty

difficulty	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
#NEs		2	2	0	0	2	0	3	9	10	12	32

Average annotation time (msec)

anno. ID	T1	T2	T1+T2	accuracy
01	4,862	2,090	6,952	0.82
02	13,115	2,998	16,114	0.75
04	7,753	1,383	9,136	0.88
08	8,198	3,373	11,571	0.79
10	11,197	3,037	14,234	0.89
11	5,759	1,734	7,493	0.92
12	6,525	2,749	9,273	0.89
13	4,257	3,105	7,362	0.65
14	9,965	2,439	12,404	0.82
15	3,580	2,067	5,647	0.92
ave.	7,521	2,498	10,019	0.832

Average annotation time (T1; sec)
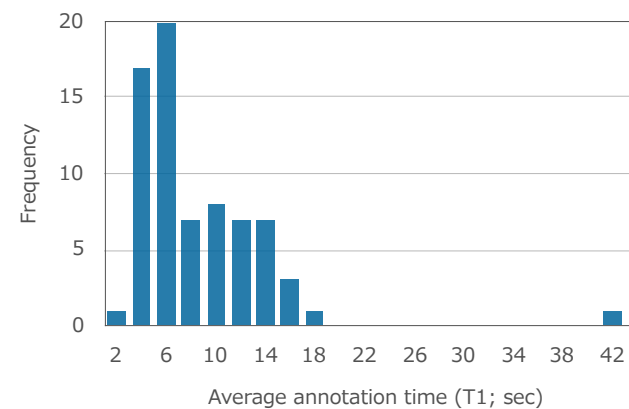
Correlation coefficient (Pearson)

annotation time vs.	T1	T2	T1+T2
difficulty	-0.47	-0.59	-0.54
text length	0.36	0.12	0.35

Eye tracking data

ミトでお世話になった女性通訳のグエン・レ・ホアさんは「南北送電線の開通で、デルタの無灯火村に電気が通じるようになりました」と中州の島を指さした。その島にある彼女の実家を訪問すると、土間にテレビが飾ってあった。昔、サイゴンの劇場の楽団にいたという父親が、古びたバイオリンのほこりを払い、突然の訪問者に、「慕情」や「上を向いて歩こう」、それにクラシックの名曲の数々を奏でてくれた。貧しいが、メコン・デルタの農民は、世界に開けていると思った。

Fixations

ミトでお世話になった女性通訳のグエン・レ・ホアさんは「南北送電線の開通で、デルタの無灯火村に電気が通じるようになりました」と中州の島を指さした。その島にある彼女の実家を訪問すると、土間にテレビが飾ってあった。昔、サイゴンの劇場の楽団にいたという父親が、古びたバイオリンのほこりを払い、突然の訪問者に、「慕情」や「上を向いて歩こう」、それにクラシックの名曲の数々を奏でてくれた。貧しいが、メコン・デルタの農民は、世界に開けていると思った。

Average fixation ratio in local contexts

- using ± 2 words is a common local context for NER (Iwakura 2011, Darwish 2013, Passos+ 2014)

window width	± 1 chunk		± 2 chunks	
type/token	type	token	type	token
fixation frequency	0.24	0.34	0.31	0.40
fixation duration	0.24	0.38	0.31	0.44

What makes correct/incorrect responses

- Fixation distribution of correct/incorrect groups
- 14 NEs with competing responses

difficulty	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
#NEs		2	2	0	0	2	0	3	9	10	12	32

Correct group (Cg) vs. Incorrect group (Ig)

- Cg tends to refer to dependency relations
  they look at predicates of the target NE or other
  arguments of the predicate (9 cases)
- local contexts work both positive and negative
  way (3 cases)
  - clue expressions vs. annotator's knowledge

東京工業大学 Tokyo Institute of Technology

後半３０分に清水市商・佐藤と一対一となるピンチを思い切りのよい飛び出しで防ぐなど、スーパーセーブを連発した。攻撃は、左サイドで一年生の中谷が、巧みなドリブルとスルーパスで攻撃のリズムを作る。後半１８分に中谷の足がつると、同じ一年生で「一発がある」東が交代出場。東は３４分、ＣＫからＧＫがパンチしたこぼれ球をけり込み、最高のタイミングで決勝点を奪った。「やれると信じていた。三年生を中心にコツコツ我慢し、怖いもの知らずの一年生の結果を出した」。上間監督の勝因分析は冷静だった。「ウチにとっては最悪、相手にとっては最高の試合でしょう。勝負は結局頑張ったほうが勝つ」と大滝監督。清水市商におごりは、なかっただろうが、初戦にすべてをかけた「挑戦者」が、気持ちの面で、優勝を意識した「本命」を完全に上回った。

東京工業大学 Tokyo Institute of Technology

後半３０分に清水市商・佐藤と一対一となるピンチを思い切りのよい飛び出しで防ぐなど、スーパーセーブを連発した。攻撃は、左サイドで一年生の中谷が、巧みなドリブルとスルーパスで攻撃のリズムを作る。後半１８分に中谷の足がつると、同じ一年生で「一発がある」東が交代出場。東は３４分、ＣＫからＧＫがパンチしたこぼれ球をけり込み、最高のタイミングで決勝点を奪った。「やれると信じていた。三年生を中心にコツコツ我慢し、怖いもの知らずの一年生の結果を出した」。上間監督の勝因分析は冷静だった。「ウチにとっては最悪、相手にとっては最高の試合でしょう。勝負は結局頑張ったほうが勝つ」と大滝監督。清水市商におごりは、なかっただろうが、初戦にすべてをかけた「挑戦者」が、気持ちの面で、優勝を意識した「本命」を完全に上回った。

東京工業大学 Tokyo Institute of Technology

サッポロビールは、コニャックの名門・仏フラパン社の最高級コニャック「フラパン・ラブレー・バカラ」＝写真＝４５本を１本５０万円で限定発売する。作家フランソワ・ラブレーの生誕５００年を記念し、フラパン社が発売する５００本の一部。平均熟成年数は５０−６０年。容器には２４金で孔雀（くじゃく）などを装飾したバカラ社特製クリスタル・デカンタを使用。アルコール分４０％で、７００ミリリットル入り。予約受け付けは２６日から来月３１日まで、はがきで「〒１５０東京都渋谷区恵比寿４の２０の１サッポロビール（株）ワイン洋酒事業部『フラパン　ラブレー・バカラ』係」。申し込みが限定数量を上回った場合は抽選する。

東京工業大学 Tokyo Institute of Technology

サッポロビールは、コニャックの名門・仏フラパン社の最高級コニャック「フラパン・ラブレー・バカラ」＝写真＝４５本を１本５０万円で限定発売する。作家フランソワ・ラブレーの生誕５００年を記念し、フラパン社が発売する５００本の一部。平均熟成年数は５０−６０年。容器には２４金で孔雀（くじゃく）などを装飾したバカラ社特製クリスタル・デカンタを使用。アルコール分４０％で、７００ミリリットル入り。予約受け付けは２６日から来月３１日まで、はがきで「〒１５０東京都渋谷区恵比寿４の２０の１サッポロビール（株）ワイン洋酒事業部『フラパン　ラブレー・バカラ』係」。申し込みが限定数量を上回った場合は抽選する。

Slide 37

Quit ‹ 3 › Text ID: 950103056

「まさか」の敗戦を喫した山梨学院大・上田誠仁監督は、予想を上回る早大の走りに脱帽するしかなかった。「区間記録を3人が連発するとは」。早大の主力3人の爆発力が、「本命」の走りを狂わせた。起爆剤となったのは、二区の渡辺。三区では主将、小林正が軽快なピッチ走法で一年生の中馬を追い、雪の富士山を正面に見る坂で一気に抜き去った。「渡辺さんの走りを見て興奮。オーバーペース気味になった」と言う四区・小林雅もそのまま押し切り、差を1分39秒に広げた。五区で山梨の主将、小椋がいったん40秒差にまで迫ったが、オーバーペース。ゴールでの差は1分47秒にまで開いていた。前回は「本命」と言われながら完敗。花田らエース級3人が卒業して迎えた今年は、山梨との立場が全く逆転した。

Slide 38

Quit ‹ 3 › Text ID: 950103056

（同じテキストに注釈の円が重ねられている）

Slide 39

Quit ‹ 26 › Text ID: 950109041

四月二十三日に第一回投票、五月七日に第二回投票が行われる、仏大統領選挙をめぐる立候補者の最終的な絞り込み作業が、年明けとともに白熱してきた。保守、中道が乱立傾向を示す一方、ドロール欧州連合委員長の不出馬で、意気消沈する左翼陣営は土壇場の対応を迫られている。同大統領選に今のところ、正式な出馬表明をしているのは、ドゴール主義の継承を掲げる保守の共和国連合のシラク元首相ら四人。各種世論調査でリードを続けたドロール氏の不出馬表明後、独走態勢にある保守・パラデュール首相は今月二十日前後に、沈黙を破り正式出馬表明を行う可能性が強くなった。先月下旬のフランス航空機乗っ取り事件の、手際の良い処理で人気を高め、八日までに、ベルナール・ドブレ経済協力相ら十三閣僚がパラデュール氏支持を表明した。社会党は二月上旬までに候補者選出の予定だが、有力候補に次々と辞退され人選作業は難航している。この間げきを縫い、反欧州統合を唱える保守「もう一つの欧州」のドビリエ党首、中道のバール元首相などが独自の立場から出馬をうかがっている。

Slide 40

Quit ‹ 26 › Text ID: 950109041

（同じテキストに注釈の円が重ねられている）

Linking task: predicate-argument structure

- Identify arguments (subject, object, ⋯) of a give predicates

kinô	*watasi*	*wa*	*tomodati*	*to*	*ranti*	*wo*	*tabemasita*
yesterday	I	NON	friend	WITH	lunch	ACC	had

Yesterday, I had lunch with my friend.

have		
subj	*watasi*	I
obj	*ranti*	lunch
dat	–	–

Intra- vs. Inter-sentential argument

intra-sentential argument

I had lunch with my friend yesterday.

φ was satisfied with the taste.

inter-sentential argument

Identifying subj case

I had lunch with my friend yesterday.

training data

$$\left.\begin{array}{l} \vec{w}\cdot\vec{x_1} \\ \vec{w}\cdot\vec{x_2} \\ \vec{w}\cdot\vec{x_3} \\ \vec{w}\cdot\vec{x_4} \end{array}\right\} \rightarrow \arg\max_i \vec{w}\cdot\vec{x_i}$$

Training data

I had lunch with my friend yesterday.

φ was satisfied with the taste.

candidates	label
I	○
yesterday	×
my friend	×
lunch	×
taste	×

Near miss

I had lunch with my friend yesterday.

φ was satisfied with the taste.

I had lunch with my friend yesterday.

φ seemed to be satisfied with the taste.

Creating training data (ranking model)

I had lunch with my friend yesterday.

φ was satisfied with the taste.

candidates	label
I	○
yesterday	×
my friend	×
lunch	×
taste	×

candidates	fix.
I	7
yesterday	1
my friend	5
lunch	2
taste	2

Collecting eye gaze data

Data stats.

- training data
  - Japanese balanced corpus BCCWJ 184 texts
  - intra-sentential subj: 107 texts
  - inter-sentential subj: 77 texts
  - eye gaze data from 7 annotators
- evaluation data
  - BCCWJ 29,519 texts
  - intra-sentential subj: 21,816
  - inter-sentential subj: 7,703

Evaluation

- models
  - binary regression (BiReg)
  - binary ranking (BiRank)
  - fixation based ranking (FixRank)
- feature sets
  - F_{base}: lexical information
  - F_{sem}: F_{base}+semantic information
  - F_{syn}: F_{base}+syntactic dependency
  - $F_{syn+sem}$: F_{syn}+F_{sem}

Result (accuracy)

model	feature	intra	inter
BiReg	base	0.56	0.04
	sem	0.48	0.06
	syn	0.58	0.03
	sem+syn	0.52	0.05
BiRank	base	0.55	0.06
	sem	0.48	0.06
	syn	0.60	0.02
	sem+syn	0.51	0.06
FixRank	base	0.55	0.02
	sem	0.49	0.02
	syn	0.63	0.02
	sem+syn	0.58	0.02

Summary

- trend in natural language processing (NLP)
  (empiricism $\rightarrow$) rationalism $\rightarrow$ empiricism $\rightarrow$?
- human behaviour in NLP $\leftarrow$ eye gaze data
  - segmentation task: NE annotation
    - dependency relations would be a key
    - local contexts are sometime harmful
    - importance of background knowledge
  - linking task: predicate argument structure
    - dependency relations are dominant