

Is a 204 cm Man Tall or Small?

Acquisition of Numerical Common Sense from the Web

Katsuma Narisawa Yotaro Watanabe

Junta Mizuno Naoaki Okazaki Kentaro Inui

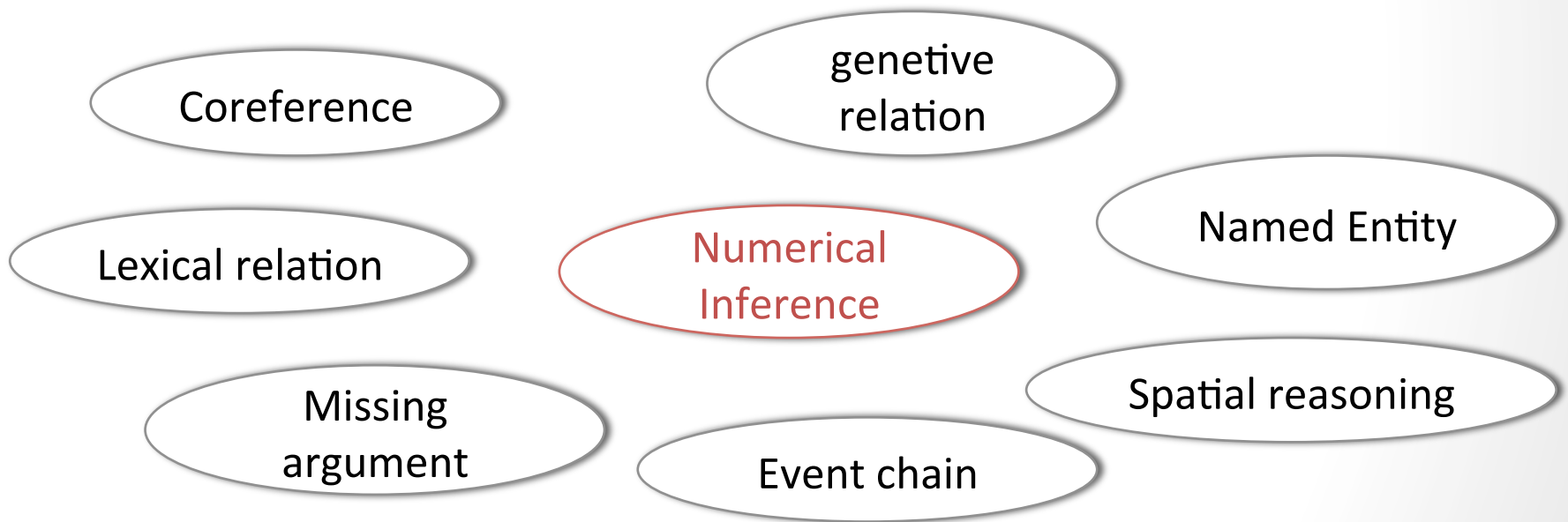
Tohoku University, Japan



TOHOKU
UNIVERSITY

Background

- Textual entailment involves many linguistic phenomena
- Time to decompose the RTE task into basic phenomena [Sammons 2010]



Even numerical inferences have variety!

Numerical matching

t: I set a target of losing two kilograms in a month.

h: I set a target of losing 2000 g in a month.

Numerical common sense

t: Before long, 3b people will face a water shortage in the world.

h: Before long, a serious water shortage will occur in the world.

Numerical
Inference

Lexical knowledge

25th wedding anniversary
⇒ silver wedding anniversary

Arithmetic

2 cats & 3 dogs ⇒ 5 animals

(We will show the distribution of these categories later)

Even numerical inferences have variety!

Numerical matching

t: I set a target of losing two kilograms in a month.
h: I set a target of losing 2000 g in a month.

Numerical common sense

t: Before long, 3b people will face a water shortage in the world.
h: Before long, a serious water shortage will occur in the world.

Numerical
Inference

This study

Lexical knowledge

25th wedding anniversary
⇒ silver wedding anniversary

Arithmetic

2 cats & 3 dogs ⇒ 5 animals

(We will show the distribution of these categories later)

We focus on *numerical common sense*

t : Before long, 3b people will face a water shortage in the world.
 h : Before long, a serious water shortage will occur in the world.

We assume that this inference is decomposable into

3b people face a water shortage

⇒ 3,000,000,000 people face (Normalize)

⇒ many people face a water shortage (Judge Large/Small)

⇒ a serious water shortage (Paraphrase?)

This study: judgment on the amount

Task definition

- To judge whether a given amount is **large**, **small**, or normal



query : He is 204 cm tall.



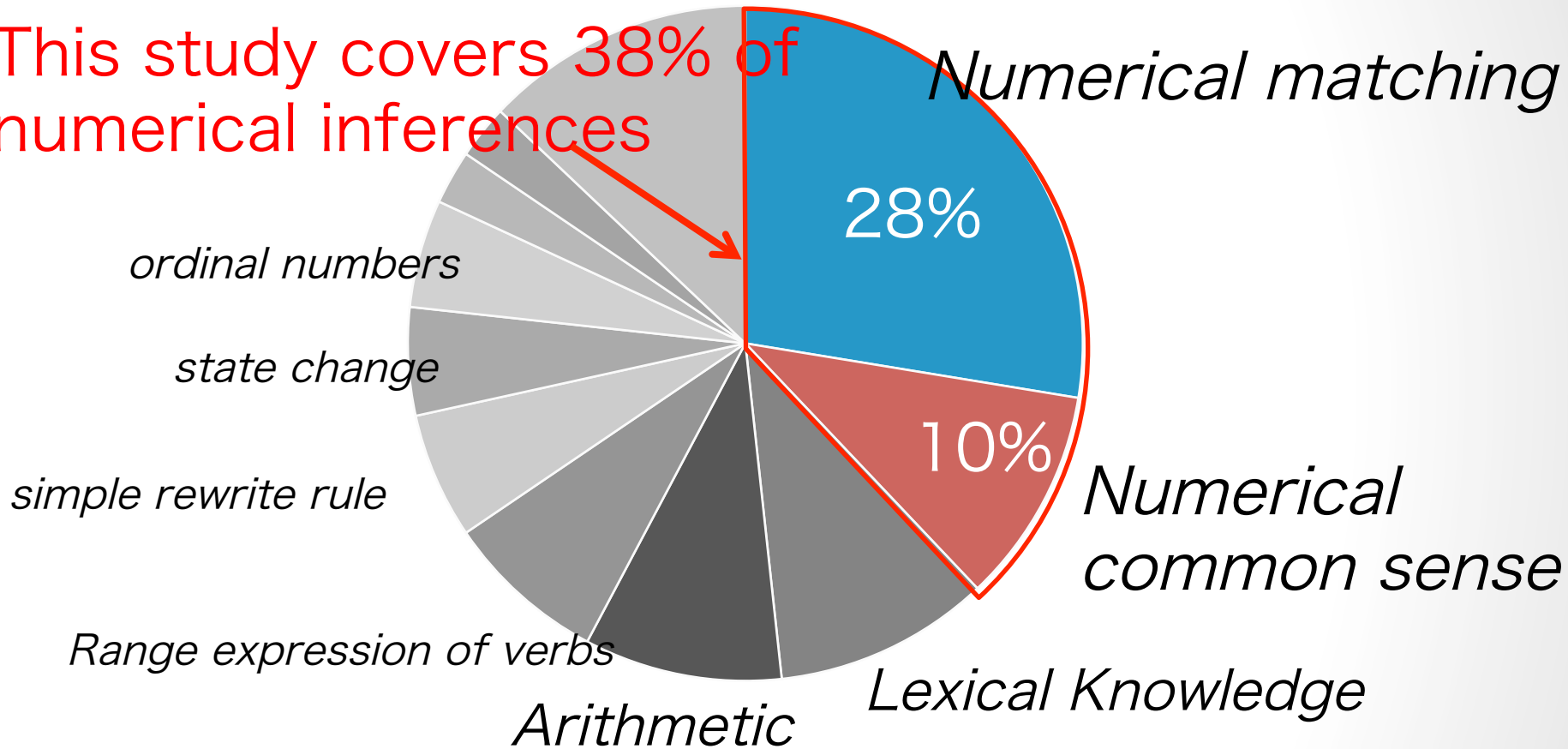
expected
output :

(As a man's height, 204cm is)

large (tall)

Textual entailment with numerical expressions

This study covers 38% of numerical inferences



114 pairs out of 4,351 TE pairs in two Japanese corpora [Shima et al 2011, Odani et al 2008] are investigated

Difficulty and solutions

- Difficulty: we need numerical common sense on various targets/situation

human's height

bird's height

price of camera

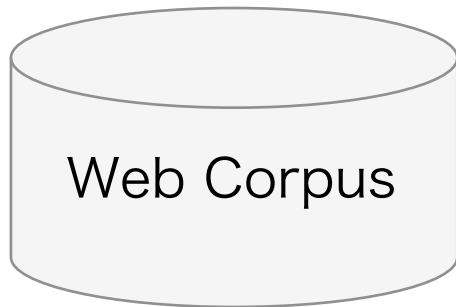
people who will face a water shortage
in the world

- Two solutions: use a large amount of texts for
 - Drawing the distribution of each target (*distribution-based*)
 - Aggregating subjective judgments on the amounts into common sense (*clue-based*)

1. Distribution-based approach

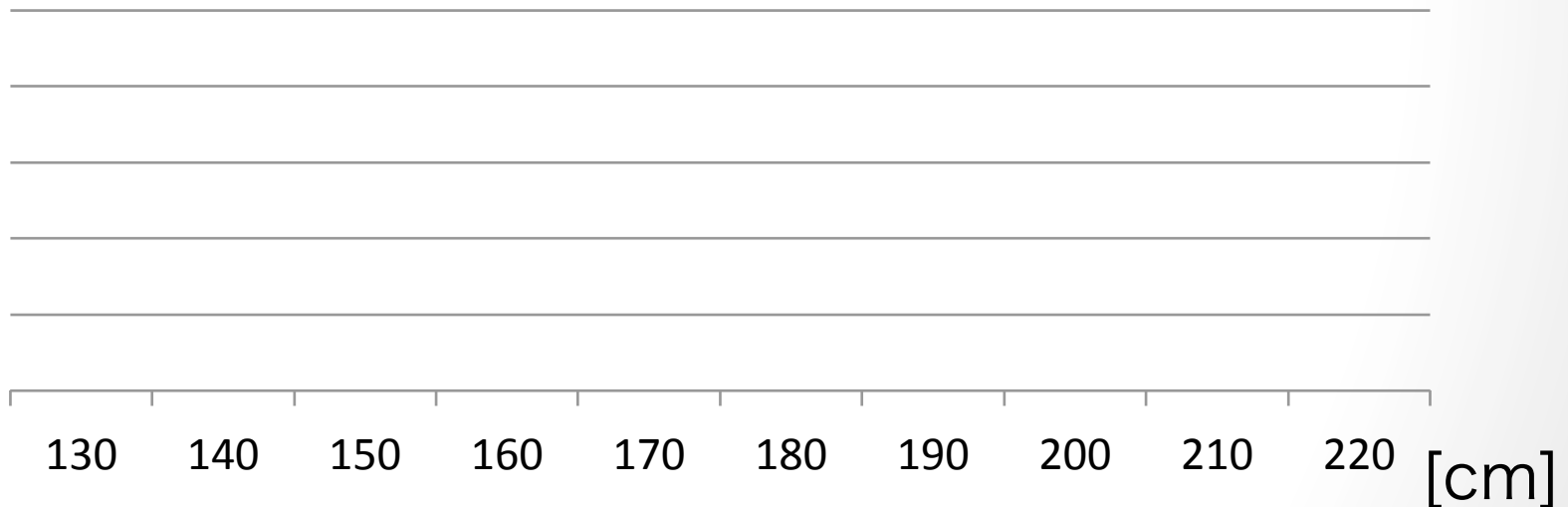
- Assumption: the real distribution of a target can be approximated by the distribution of numbers co-occurring with the target on the Web

1. Distribution-based approach

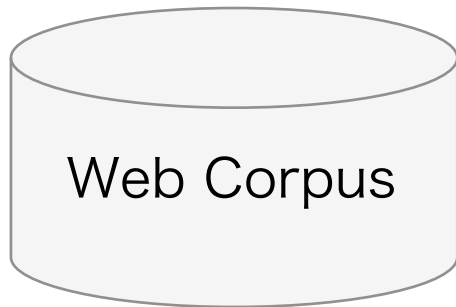


... is a Japanese actor born in Kagoshima Prefecture. **He is 170 cm tall** and weighs 62 kg. He specializes ...

men's height

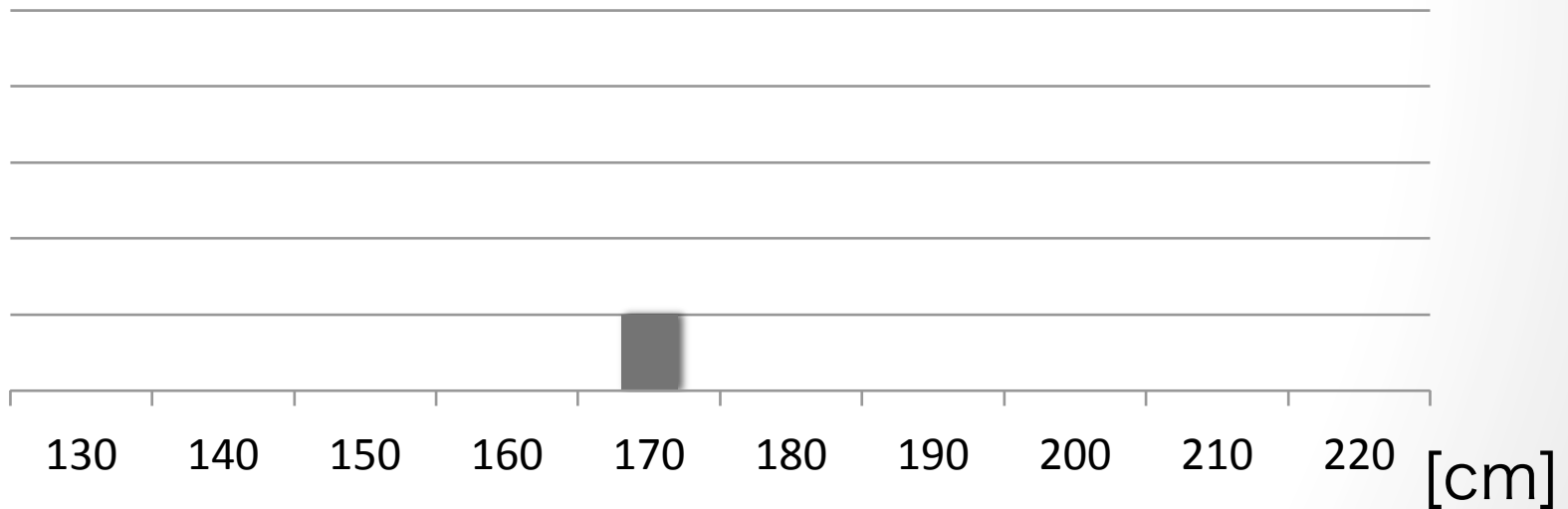


1. Distribution-based approach

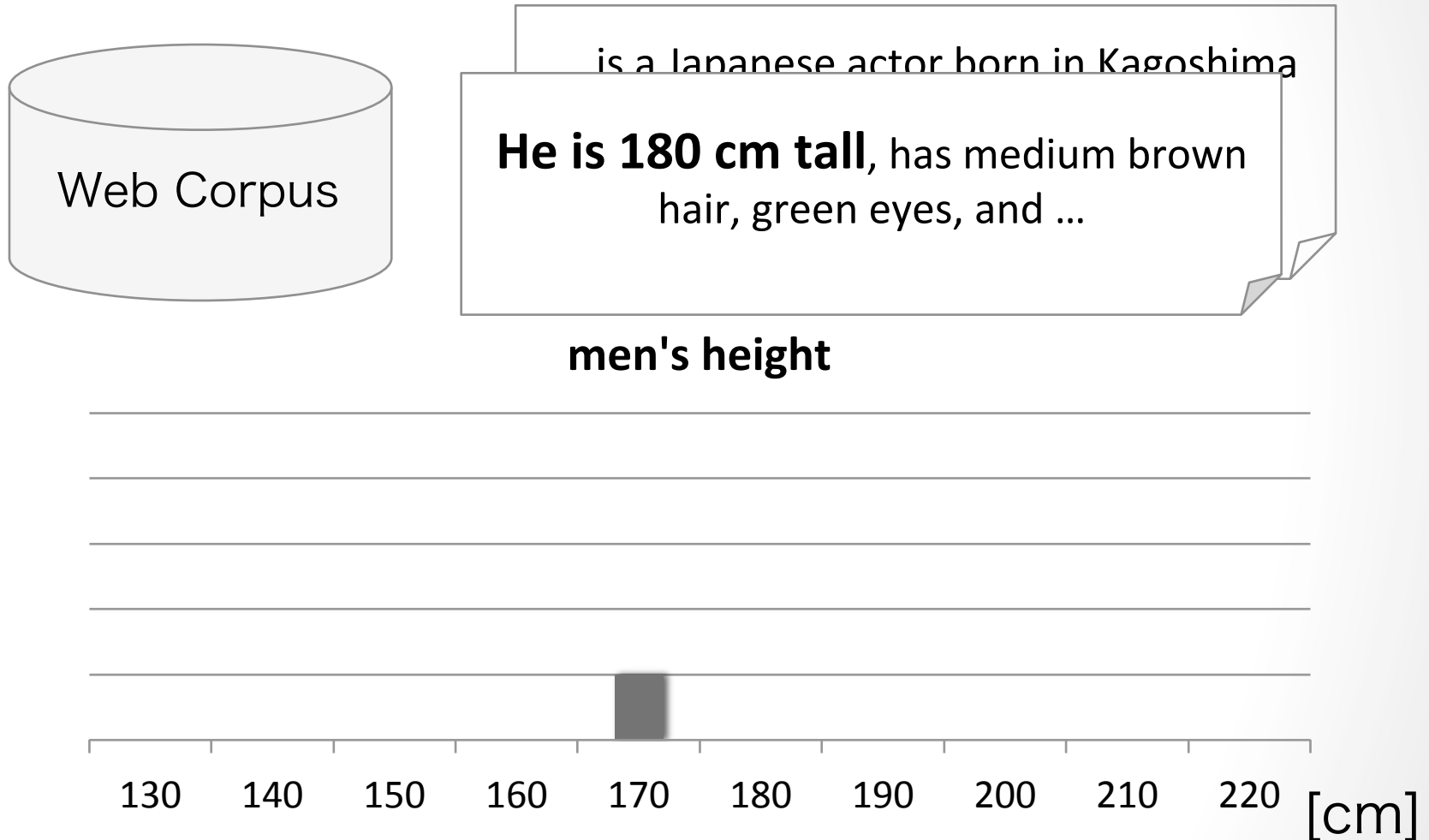


... is a Japanese actor born in Kagoshima Prefecture. **He is 170 cm tall** and weighs 62 kg. He specializes ...

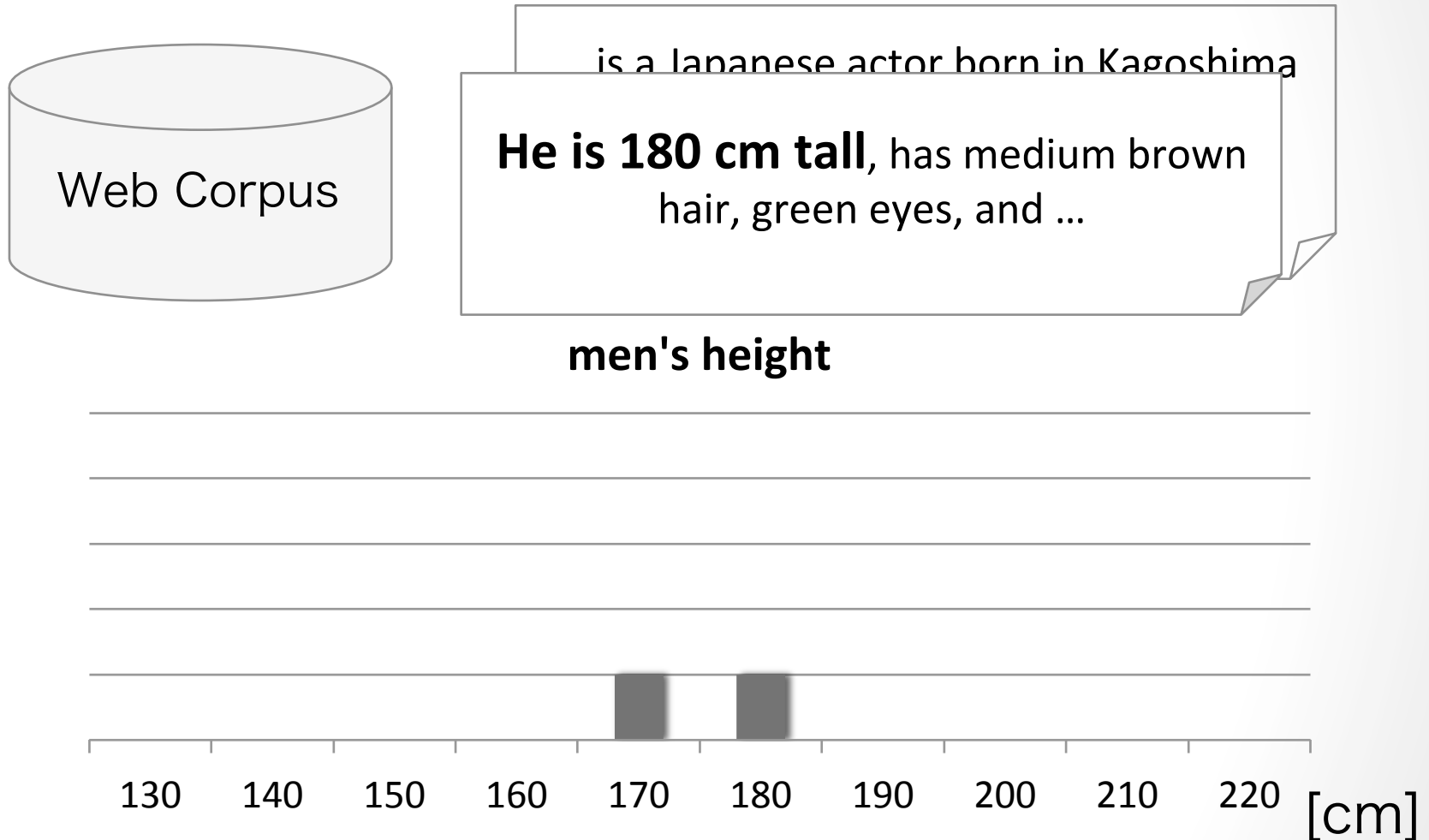
men's height



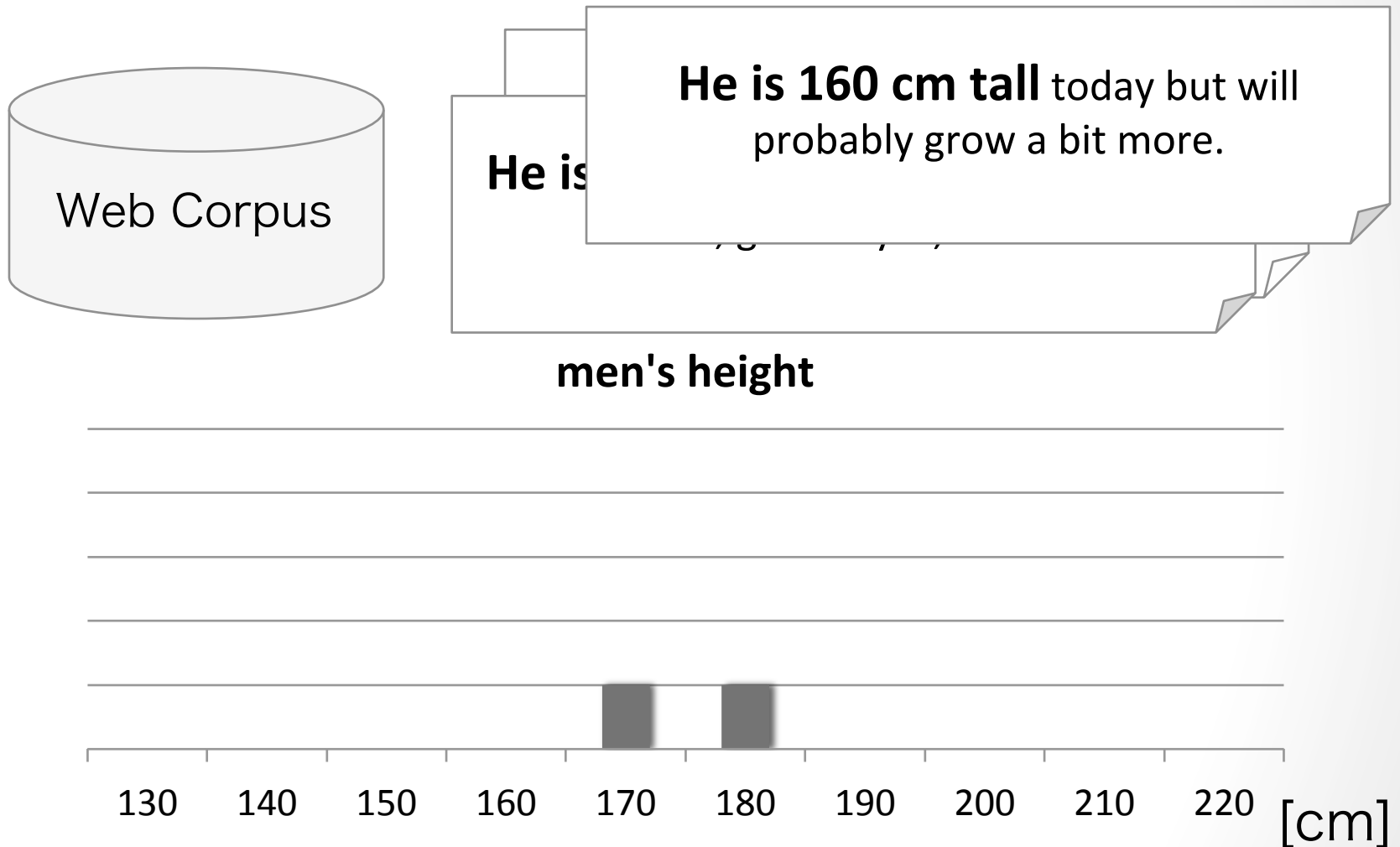
1. Distribution-based approach



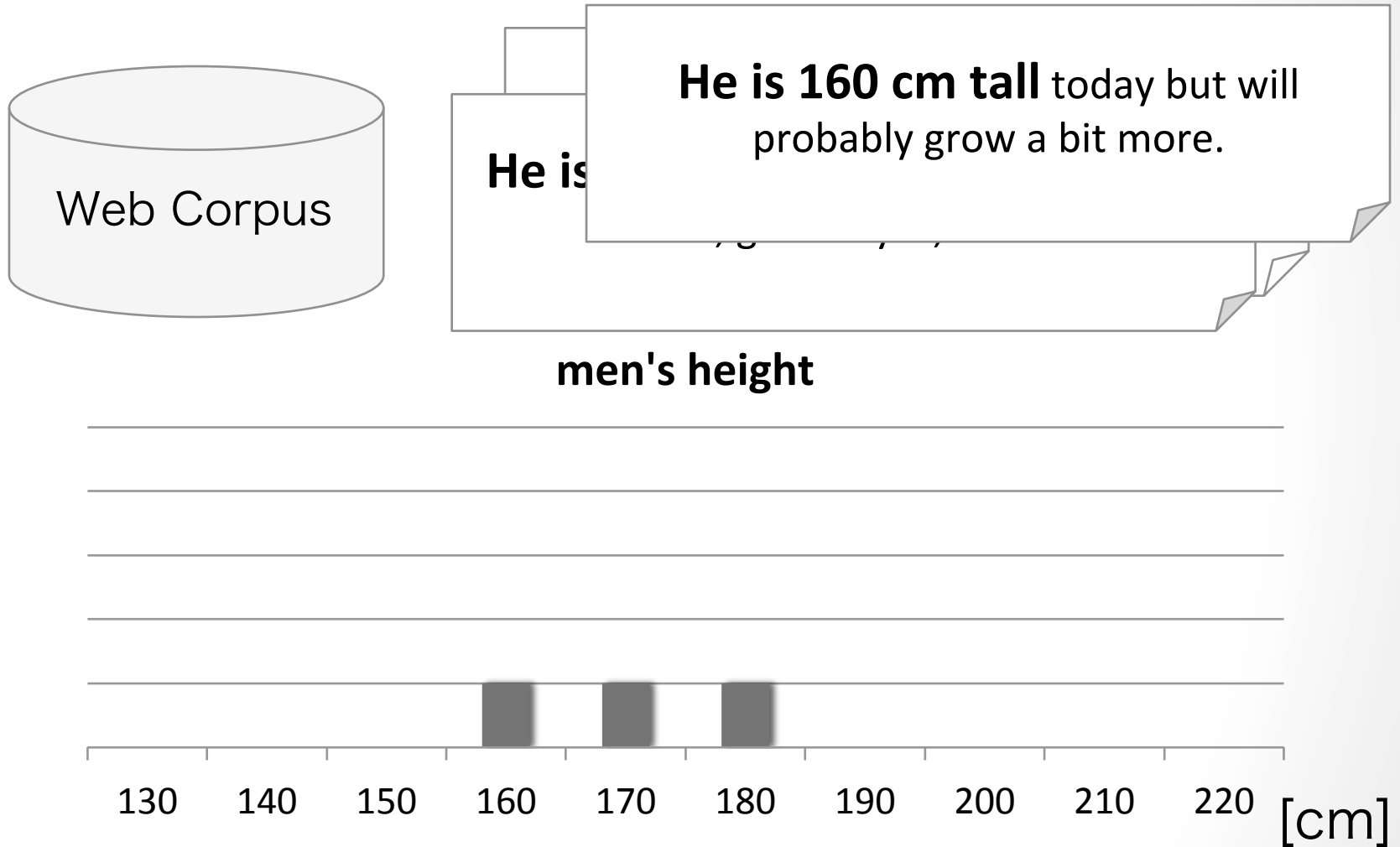
1. Distribution-based approach



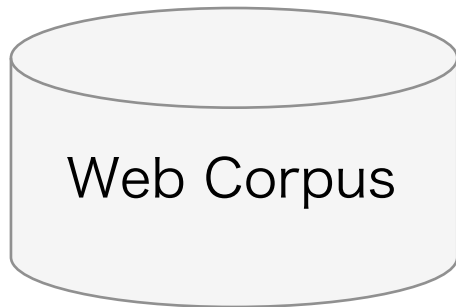
1. Distribution-based approach



1. Distribution-based approach

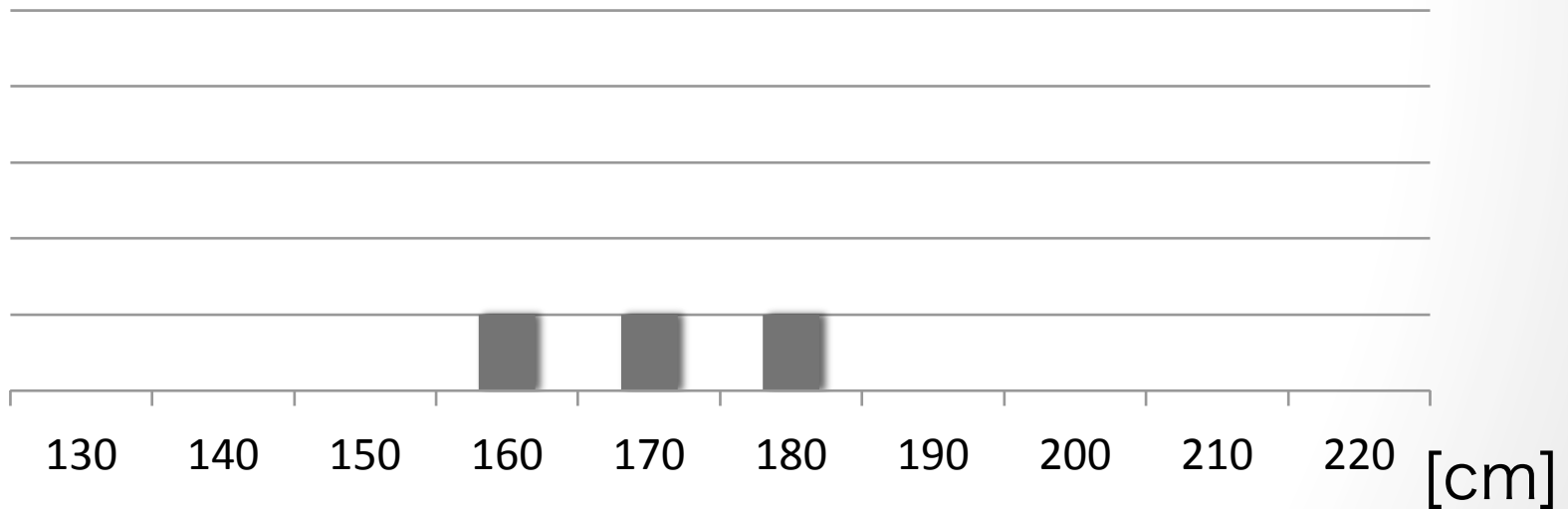


1. Distribution-based approach

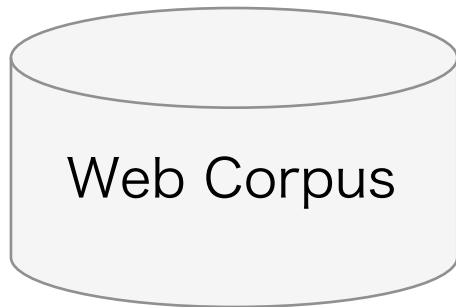


He is 160 cm tall today but will probably grow a bit more.
We do not mention too tall (or too small) men so often

men's height



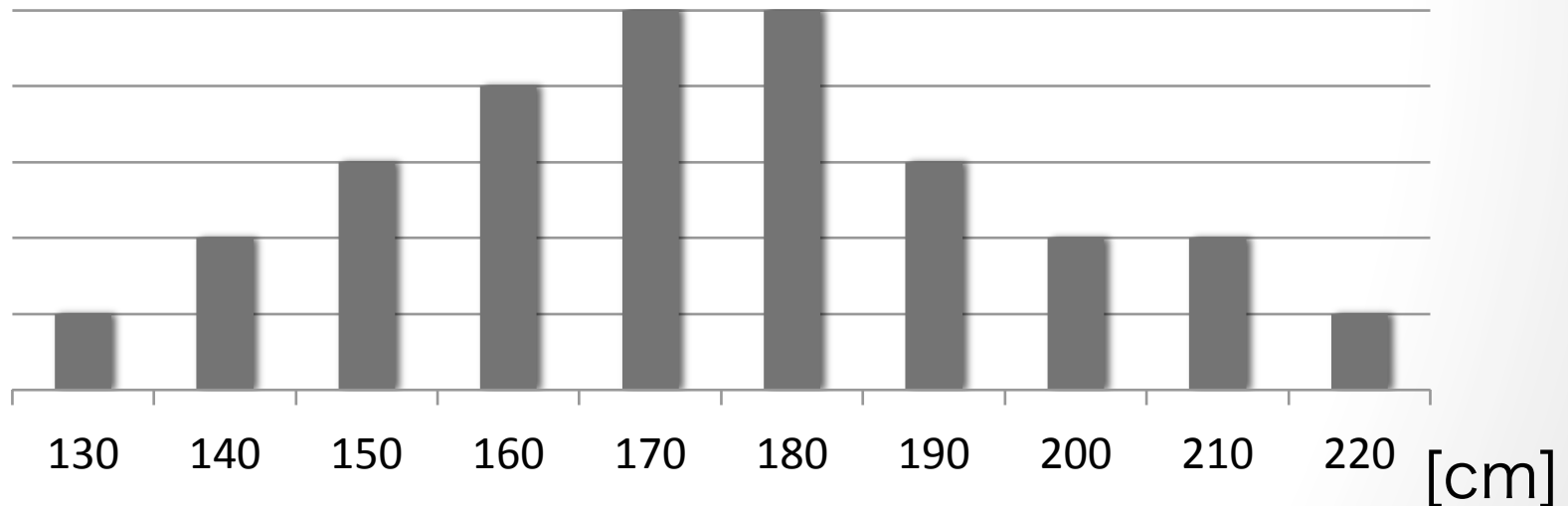
1. Distribution-based approach



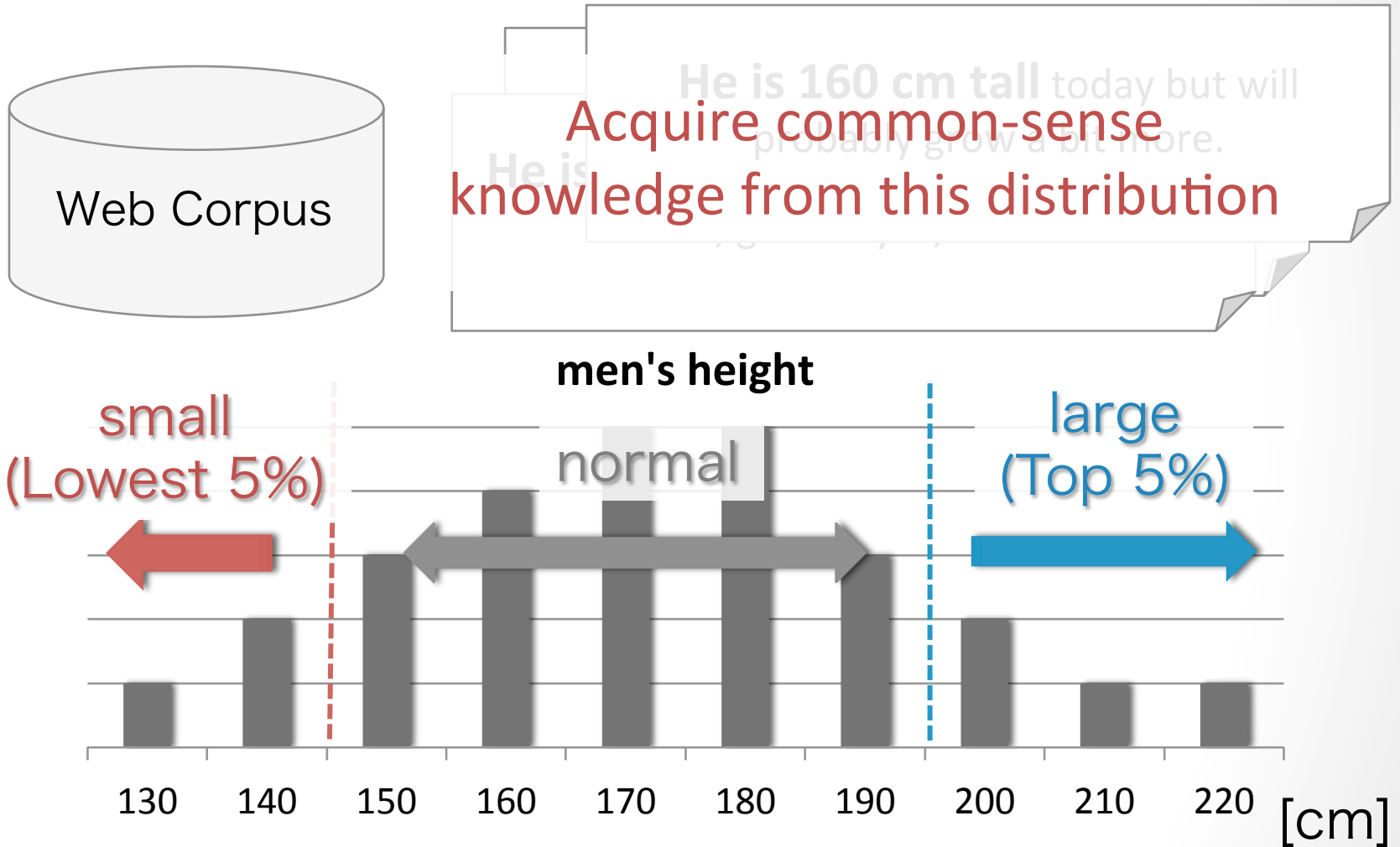
He is 160 cm tall today but will probably grow a bit more.

We do not mention too tall (or too small) men so often

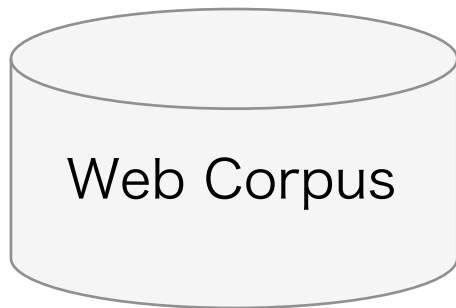
men's height



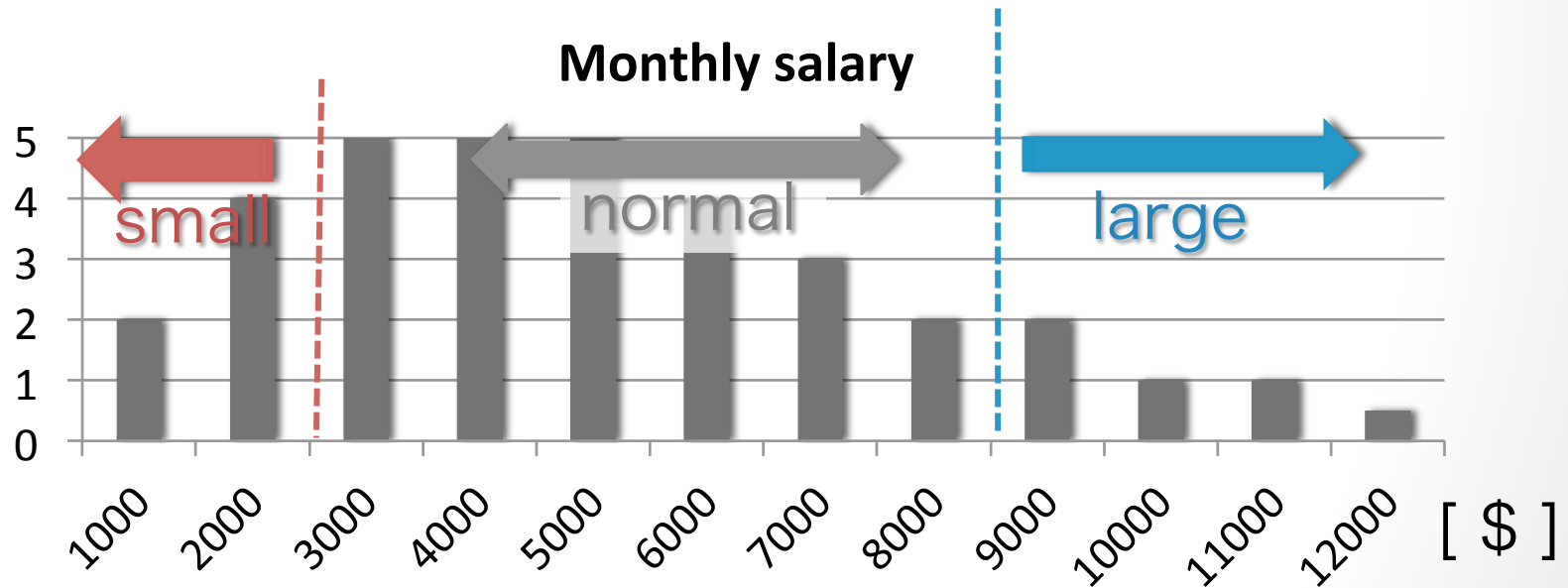
1. Distribution-based approach



1. Distribution-based approach



I earn **\$2.000 a month** and enjoy **th and**
m I am a manager of a store and I
earn \$10,000 a month on profit. ...

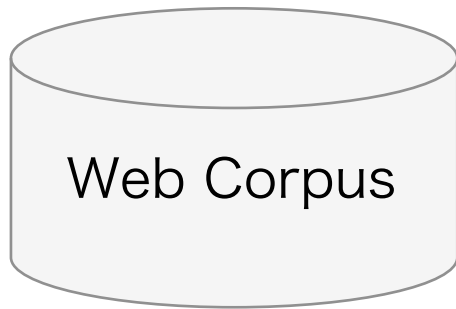


2. Clue-based approach

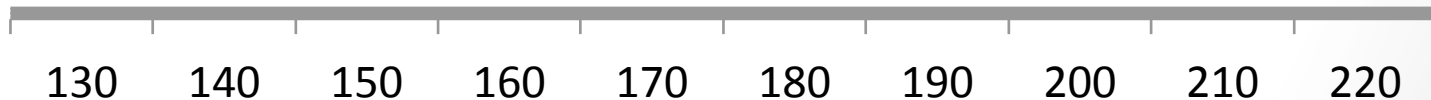
- Uses textual clues expressing humans' judgements (e.g., *only*, *as many as*)

2. Clue-based approach

- Uses textual clues expressing humans' judgements (e.g., *only, as many as*)

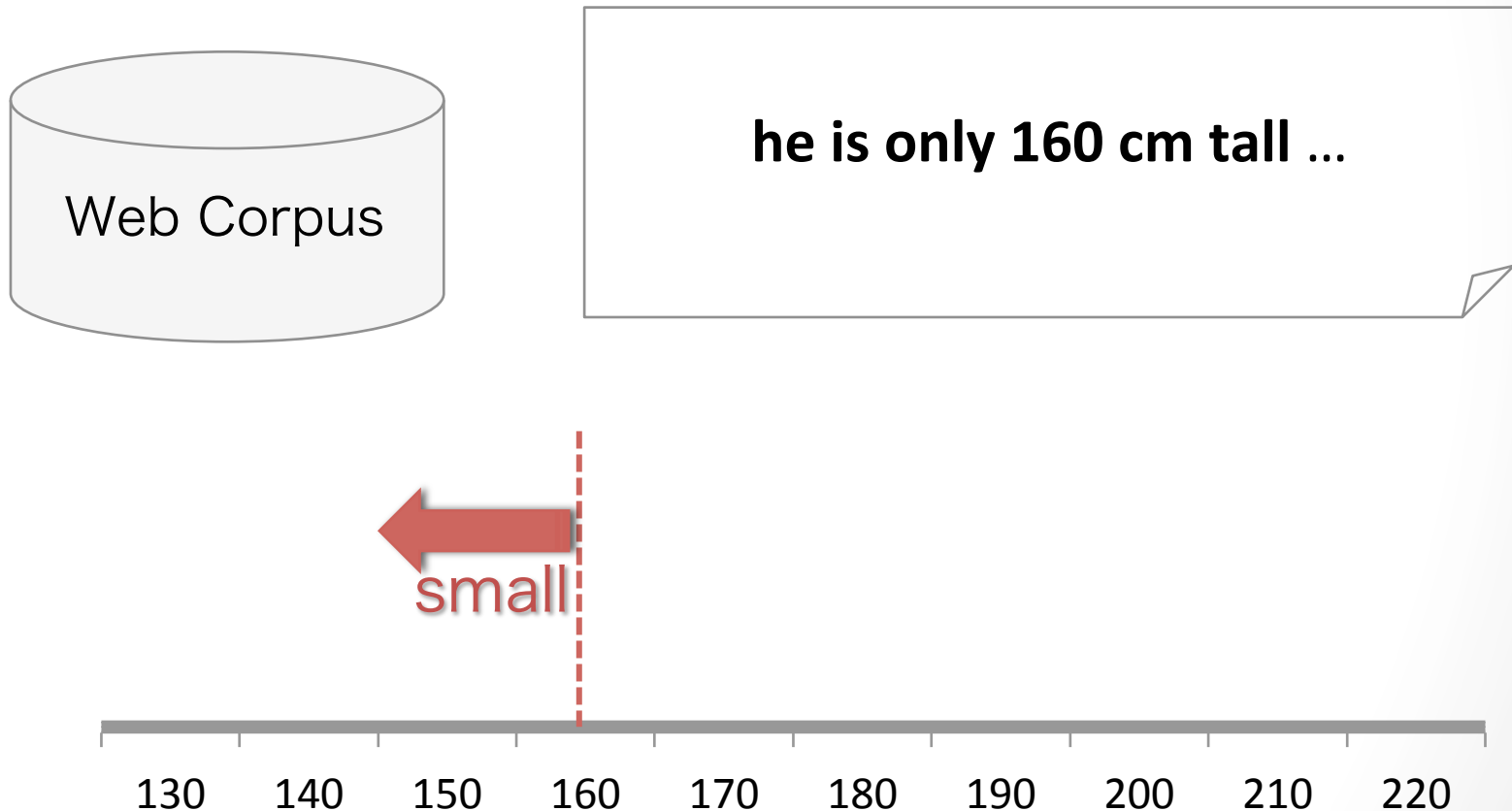


he is only 160 cm tall ...



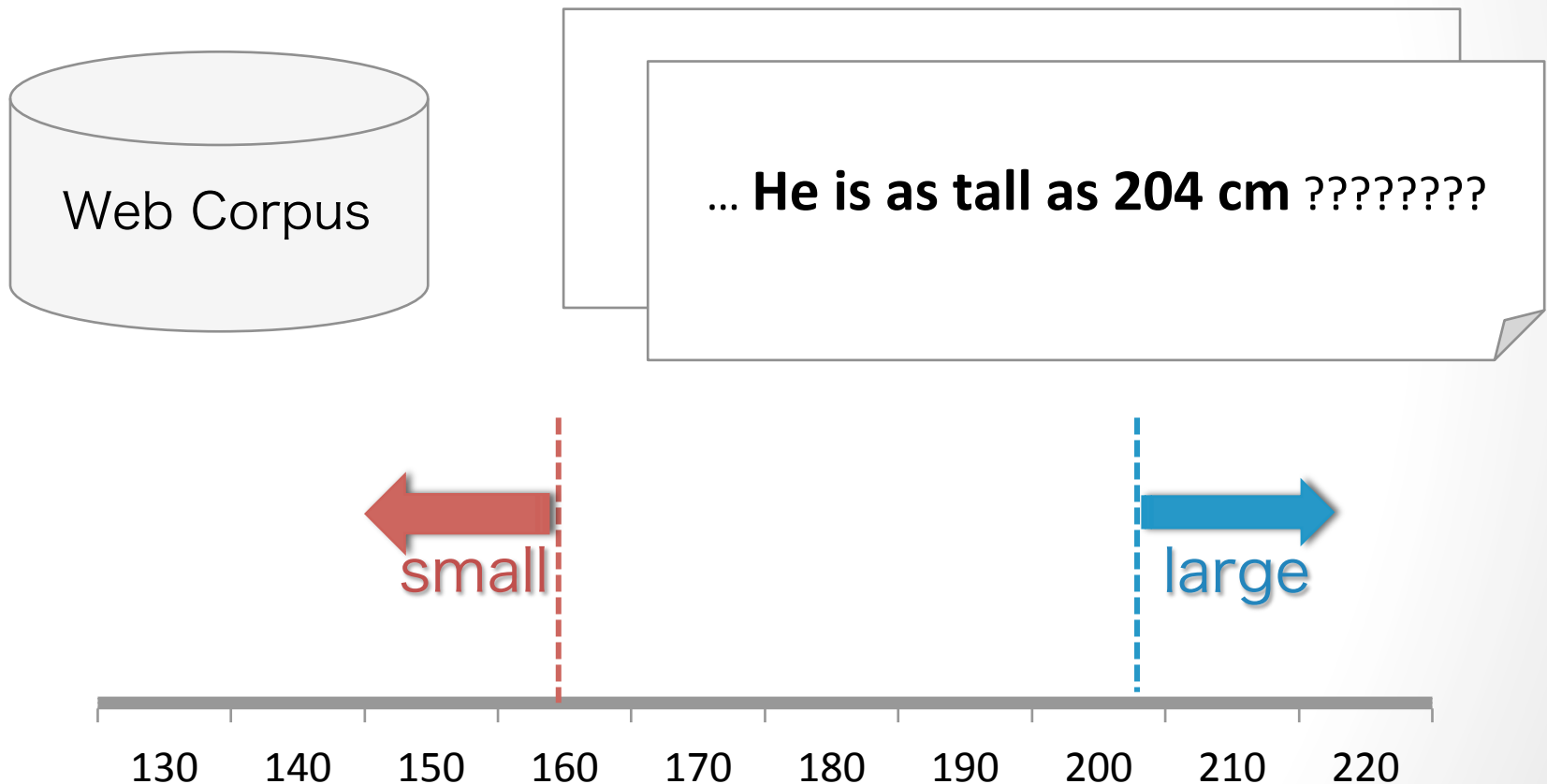
2. Clue-based approach

- Uses textual clues expressing humans' judgements (e.g., *only, as many as*)



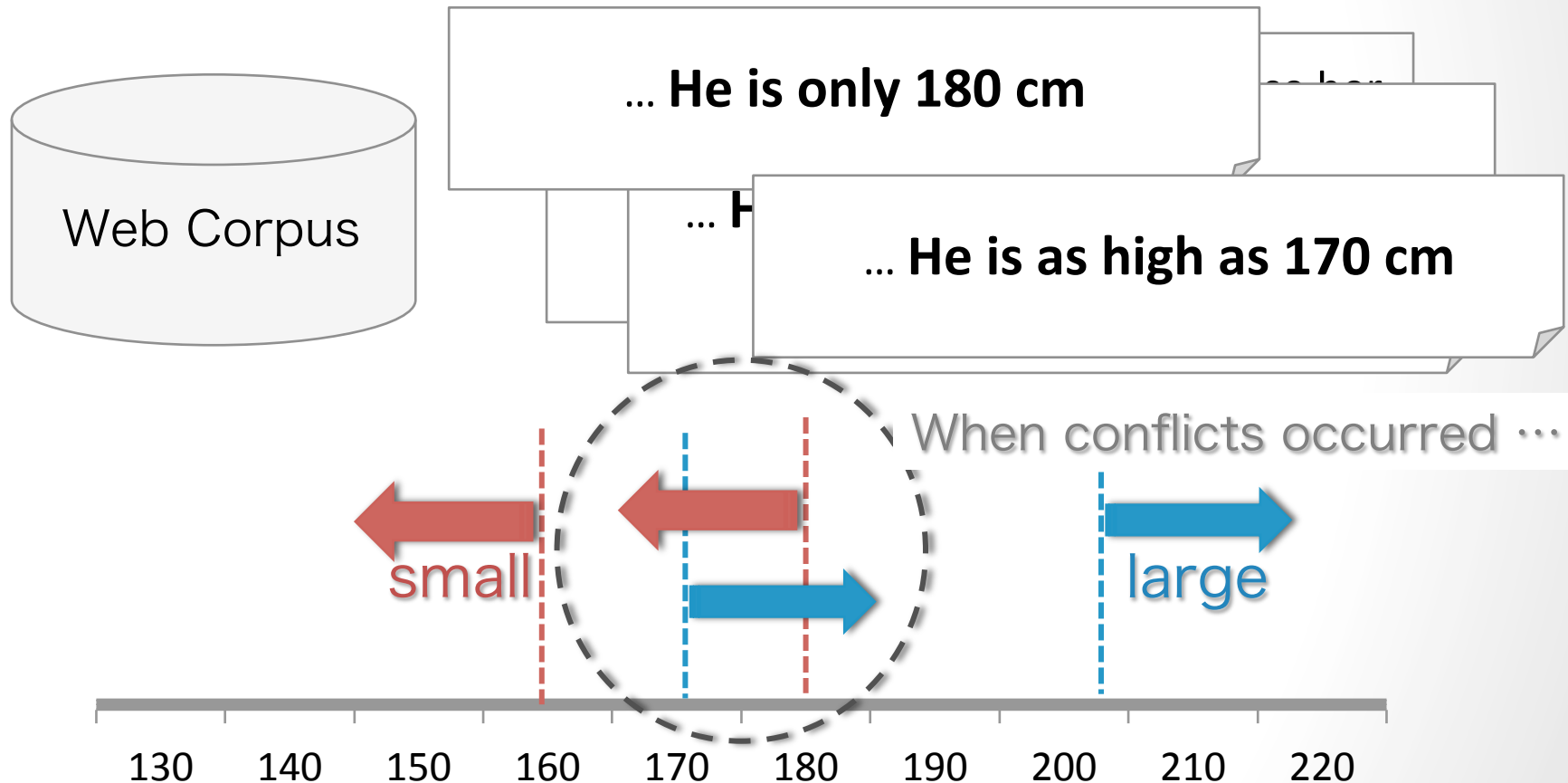
2. Clue-based approach

- Uses textual clues expressing humans' judgements (e.g., *only*, *as many as*)



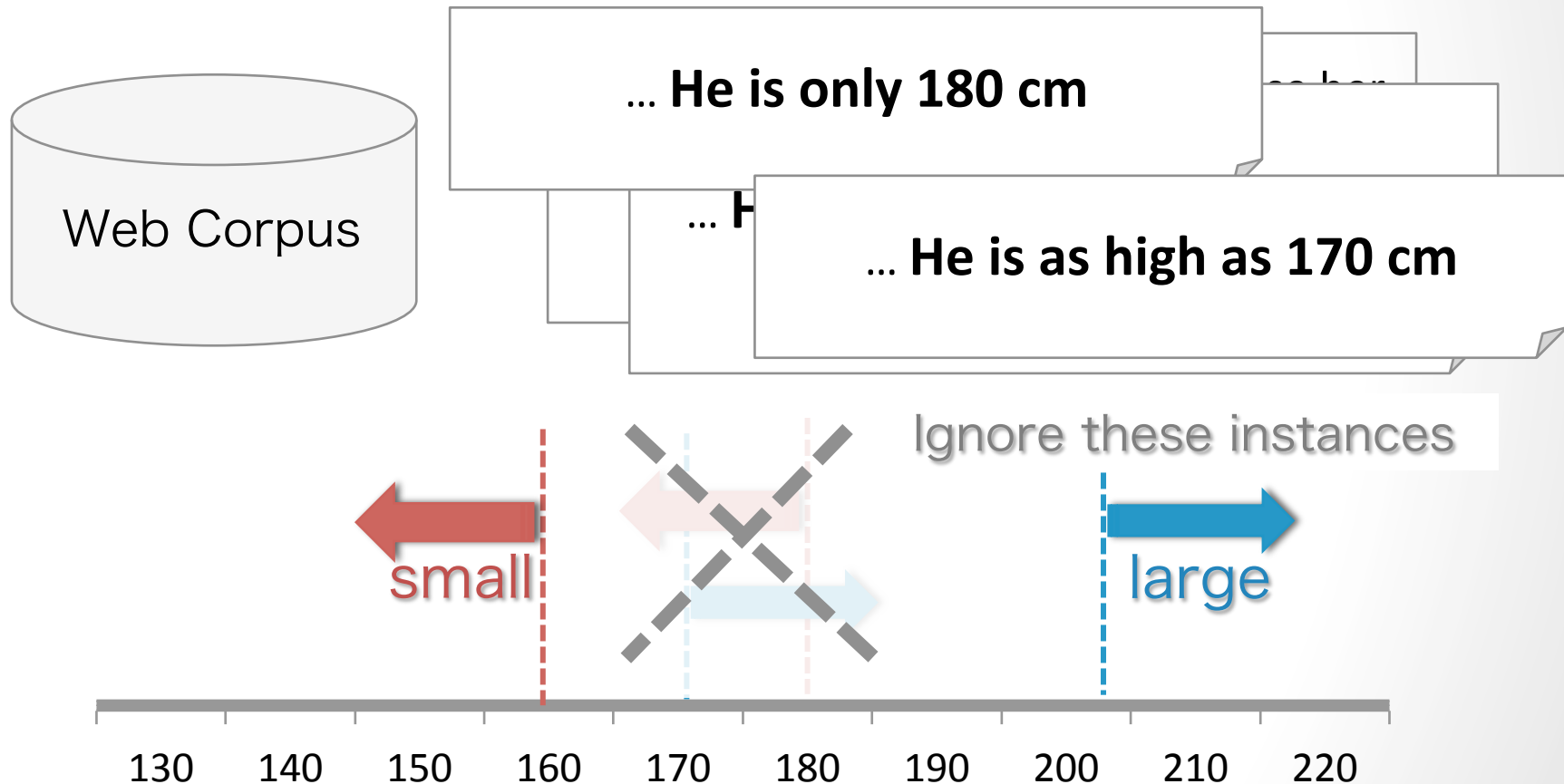
2. Clue-based approach

- Uses textual clues expressing humans' judgements (e.g., *only*, *as many as*)



2. Clue-based approach

- Uses textual clues expressing humans' judgements (e.g., *only*, *as many as*)



Review of task definition

- To judge whether a given amount is **large**, **small**, or normal



query : He is 204 cm tall.



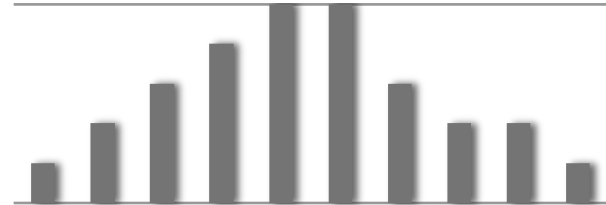
expected
output :

(As a man's height, 204cm is)

large (tall)

Review of two approaches

- Use a large amount of texts for ...
 - Distribution-based : Drawing the distribution of each target



- Clue-based : Aggregating subjective judgments on the amounts into common sense

... He is **only** 160 cm tall



Recognizing numerical expressions in text

- Both of the approaches require:
 - A. To extract and normalize numerical expressions
 - B. To extract the context of numerical expressions

A. Extract & normalize numerical expressions

input sentences

They weigh **two kilograms**.



extract & normalize

Extracted Numerical Expression	Semantic representation		
	Value	Unit	Modifier
two kilograms	2000	g	

A. Extract & normalize numerical expressions

input sentences

They weigh **two kilograms**.
They weigh **2 kg**.



extract & normalize

Extracted Numerical Expression	Semantic representation		
	Value	Unit	Modifier
two kilograms	2000	g	
2 kg	2000	g	

identify surface variation

Extract & normalize numerical expressions

input sentences

They weigh two kilograms.
They weigh 2 kg.
They weigh 2 tons.



extract & normalize

Extracted Numerical Expression	Semantic representation		
	Value	Unit	Modifier
two kilograms	2000	g	
2 kg	2000	g	
2 tons	2000000	g	

convert to
canonical units

A. Extract & normalize numerical expressions

input sentences

There are about 10 apples.



extract & normalize

recognize modifier
(e.g., about, over
more than, etc)

Extracted Numerical Expression	Semantic representation		
	Value	Unit	Modifier
about 10 apples	10	apples	about

A. Extract & normalize numerical expressions

input sentences

There are **about 10 apples**.
There are **as many as 10 apples**.
There are **only 10 apples**.



extract & normalize

recognize modifier
(e.g., about, over
more than, etc)

Extracted Numerical Expression	Semantic representation		
	Value	Unit	Modifier
about 10 apples	10	apples	about
as many as 10 apples	10	apples	large
only 10 apples	10	apples	small

A. Extract & normalize numerical expressions

- We used rule-based approach

They weigh about two kilograms.

Extracted Numerical Expression	Semantic representation		
	Value	Unit	Modifier

A. Extract & normalize numerical expressions

- We used rule-based approach

They weigh about **two** kilograms.

Extracted Numerical Expression	Semantic representation		
	Value	Unit	Modifier
two	2		

1. Find numbers in the text

A. Extract & normalize numerical expressions

- We used rule-based approach

They weigh about **two** kilograms.

Extracted Numerical Expression	Semantic representation		
	Value	Unit	Modifier
two	2		

2. Check units by using dictionary

(String)

kilogram(s)

apple(s)

...

(Operation)

set-unit: 'g'; multiply-value: 1000

set-unit: 'apples';

A. Extract & normalize numerical expressions

- We used rule-based approach

They weigh about **two kilograms**.

Extracted Numerical Expression	Semantic representation		
	Value	Unit	Modifier
two kilograms	2000	g	

2. Check units by using dictionary

(String)

kilogram(s)

apple(s)

...

(Operation)

set-unit: 'g'; multiply-value: 1000

set-unit: 'apples';

A. Extract & normalize numerical expressions

- We used rule-based approach

They weigh about **two kilograms**.

Extracted Numerical Expression	Semantic representation		
	Value	Unit	Modifier
two kilograms	2000	g	

3. Check modifiers by using dictionary

(String)
about
only
...

(Operation)
set-modifier: 'about'
set-modifier: 'small'

A. Extract & normalize numerical expressions

- We used rule-based approach

They weigh **about two kilograms**.

Extracted Numerical Expression	Semantic representation		
	Value	Unit	Modifier
about two kilograms	2000	g	about

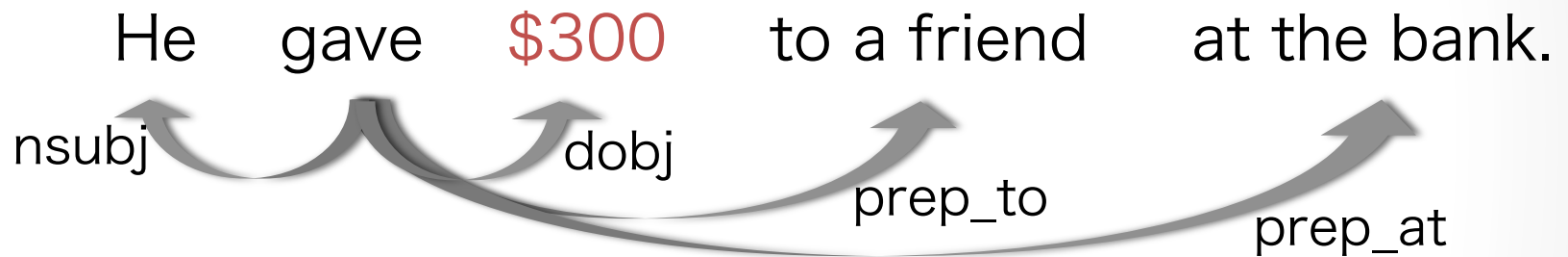
3. Check modifiers by using dictionary

(String)
about
only
...

(Operation)
set-modifier: 'about'
set-modifier: 'small'

B. Extract context of numerical expressions

- Context is defined as the verb that governs the numerical expression and its typed arguments

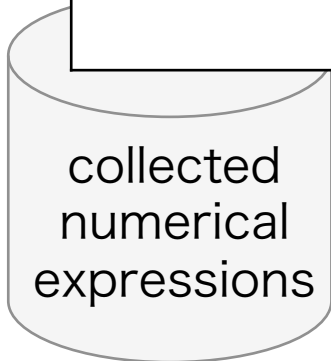
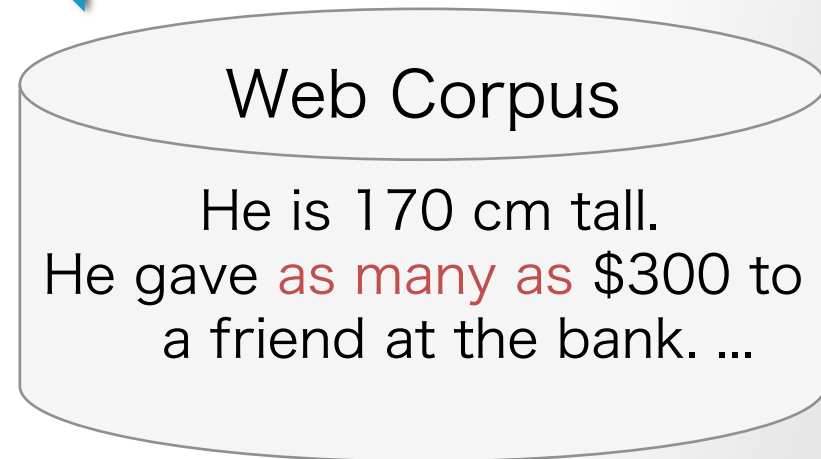


<i>context</i>	verb : <i>give</i> nsubj : <i>he</i> prep_to : <i>friend</i> prep_at : <i>bank</i>
----------------	---------------------------------------------------------------------------------------

Outline of our approach

- 0. Collect numerical expressions and their contexts from the Web

normalized expression	context
1.7, m	verb : is nsubj : he
300, \$, large	verb : give nsubj : he prep_to : friend prep_at : bank



Outline of our approach

- 1. Extract numerical expressions and context from the query

query : He is 204 cm tall

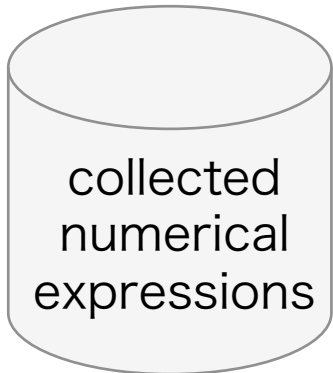
large ?

small ?

normal ?



normalized expression	context
2.04, m	verb : is nsubj : he



Outline of our approach

- 2. Retrieve numerical expressions from collected expressions which have same unit and context

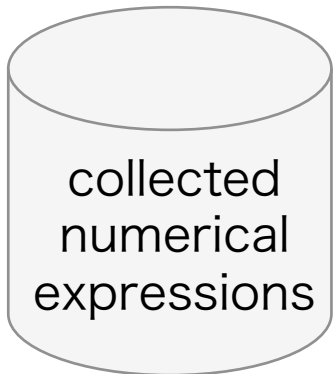
query : He is 204 cm tall

normal ?

normal ?



normalized expression	context
2.04, m	verb : is nsubj : he



1.7, m	verb : is nsubj : he
1.8, m	verb : is nsubj : he
1.3, m (small)	verb : is nsubj : he
...	...

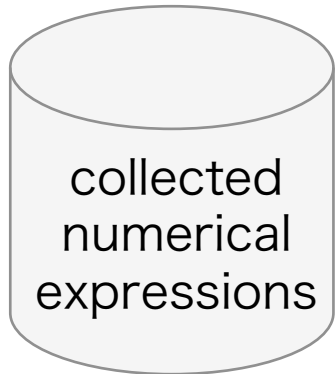
Outline of our approach

- 3. Judge by using Distribution-based / Clue-based approach

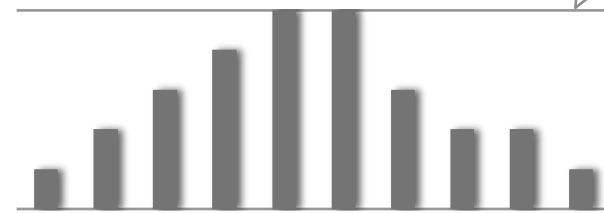
query : He is 204 cm tall



normalized expression	context
2.04, m	verb : is nsubj : he

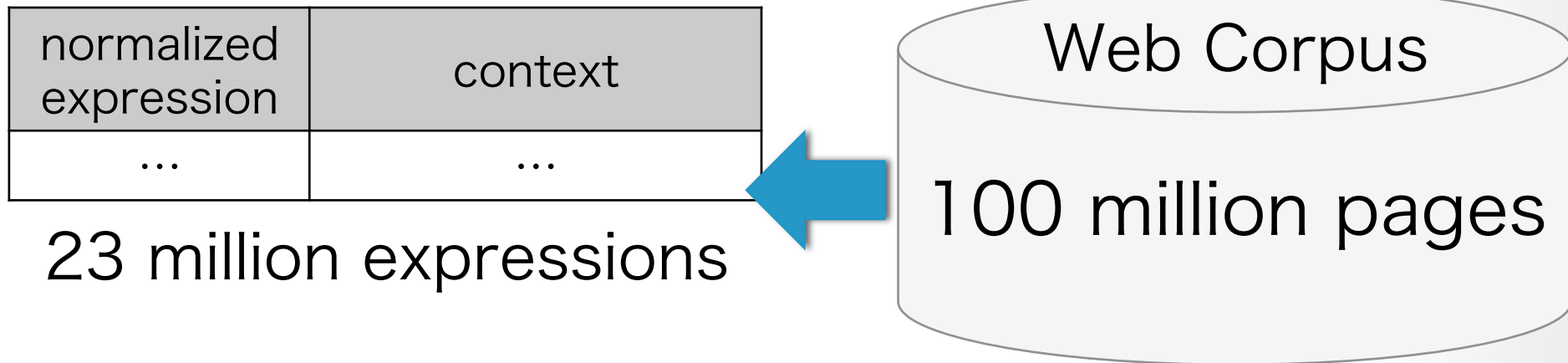


1.7, m	verb : is nsubj
1.8, m	verb : is nsubj
1.3, m (small)	verb : is nsubj
...	...



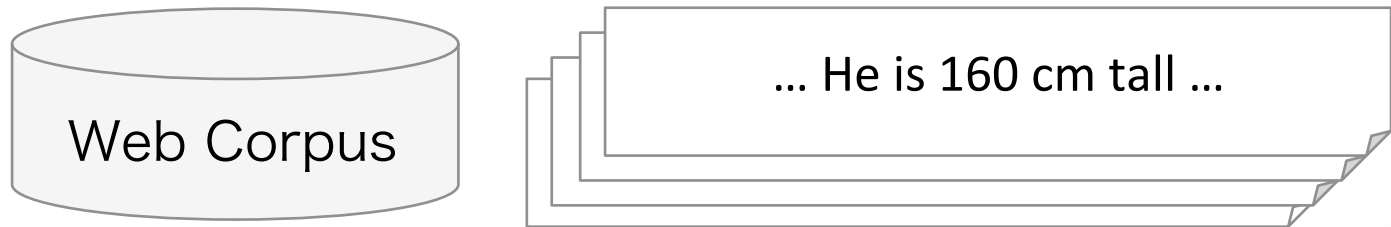
Experimental setup

- Extracted about 23 million numerical expressions from 100 million web pages



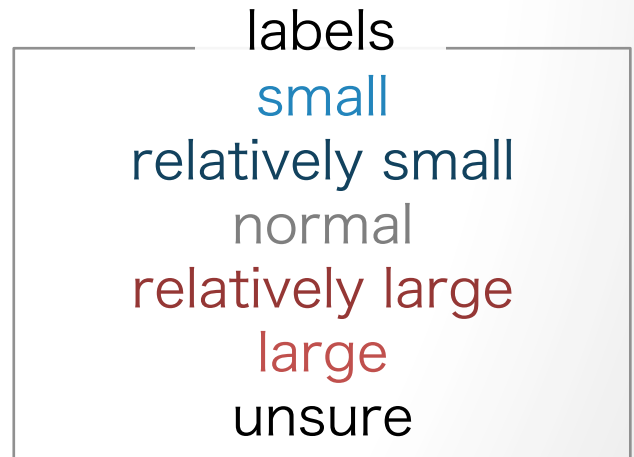
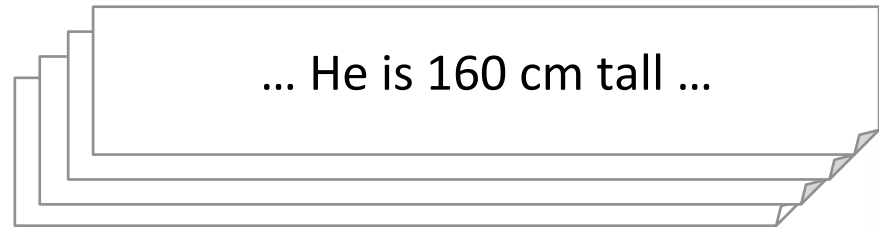
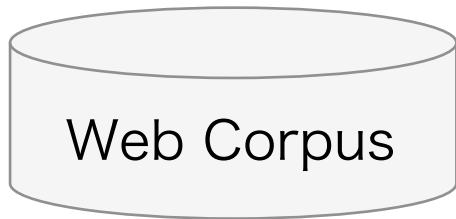
Experimental setup

- Built a gold-standard data set
 - ① We extracted 2000 sentences with numerical expressions



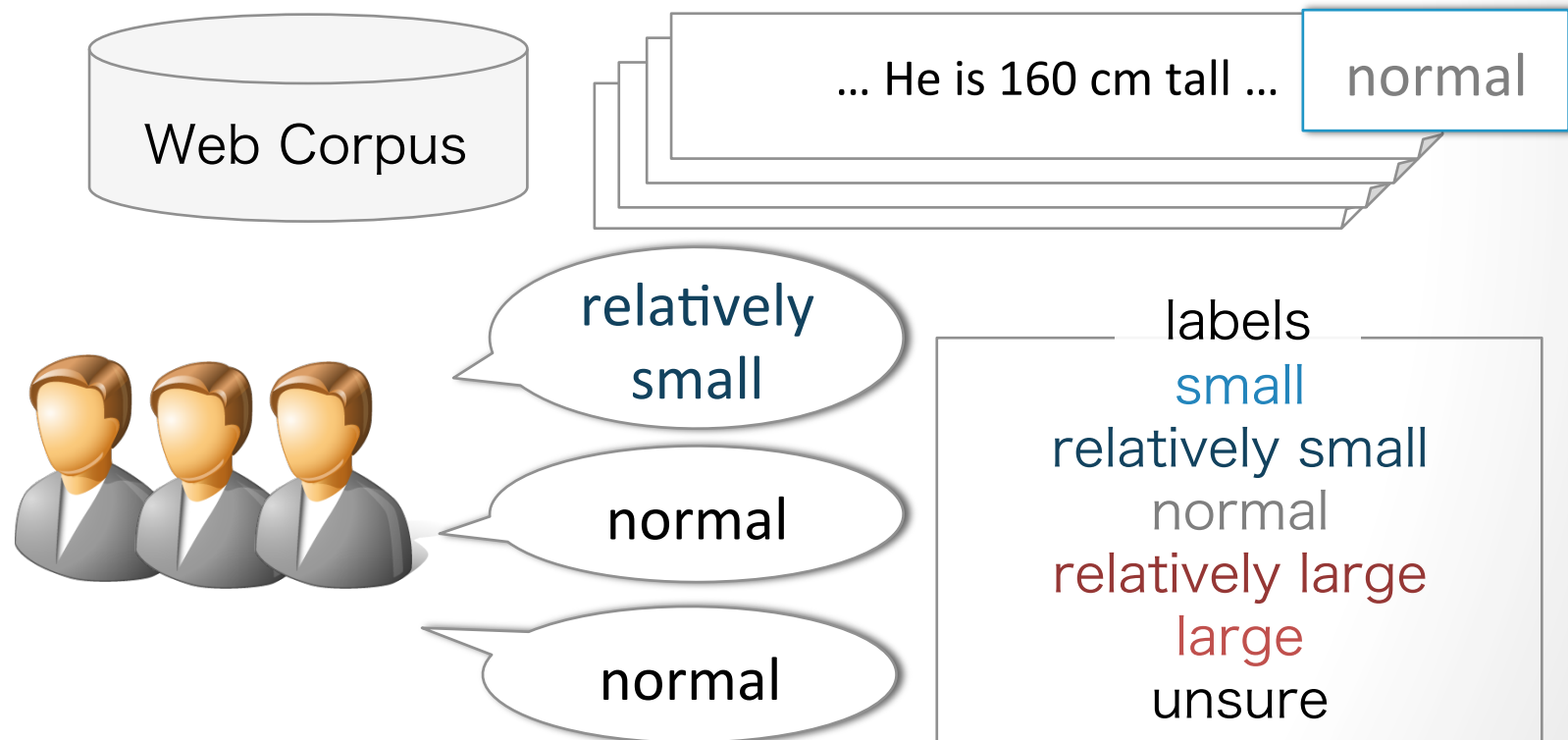
Experimental setup

- We built a gold-standard data set
 - ② We asked 3 human judges to annotate every numerical expression with one of the 6 labels



Experimental setup

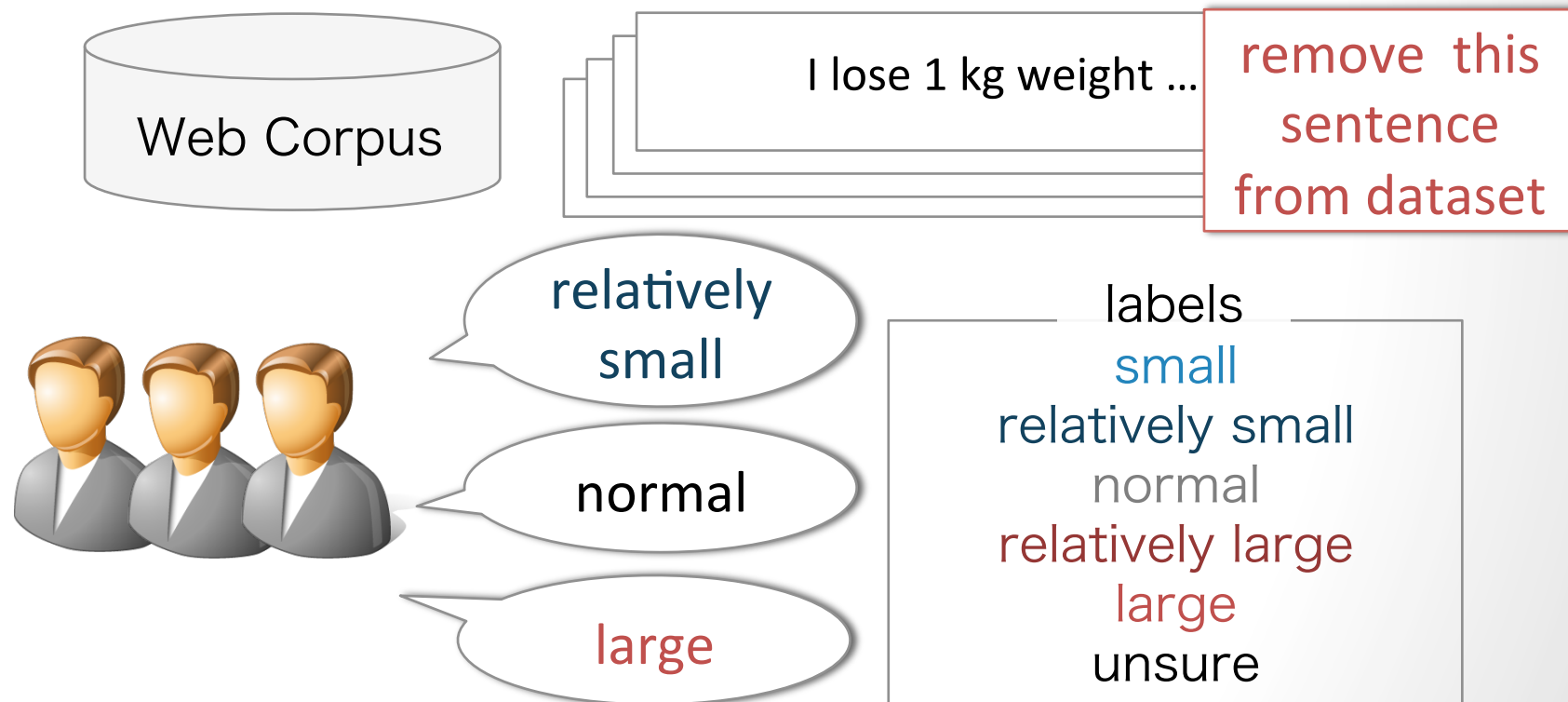
- We built a gold-standard data set
 - ③ If at least two of the annotators annotated the same label, we assigned the label to the numerical expression



Experimental setup

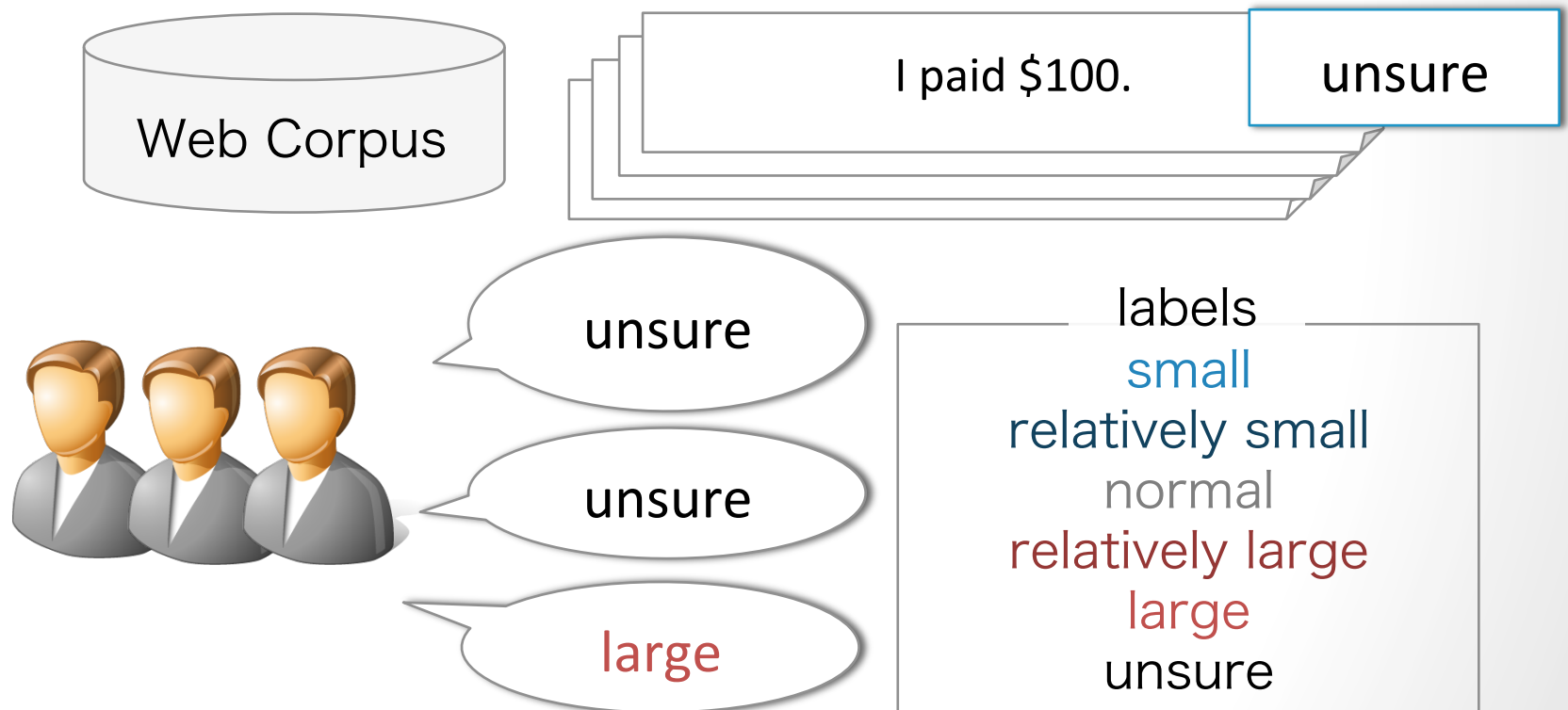
- We built a gold-standard data set

③ If at least two of the annotators annotated the same label, we assigned the label to the numerical expression



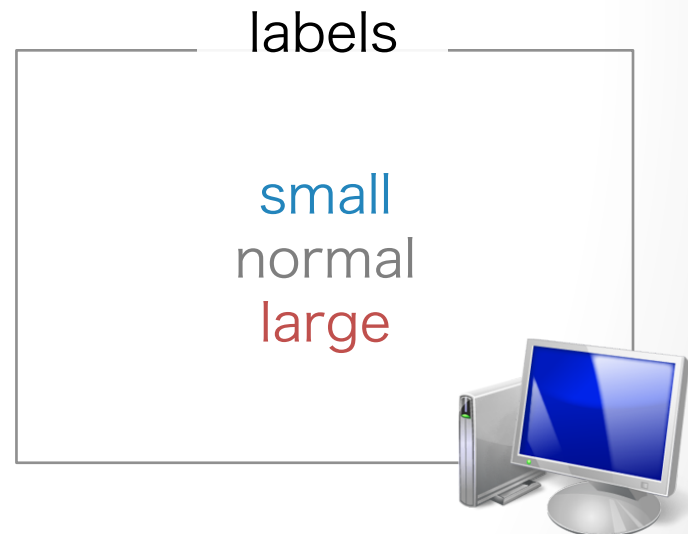
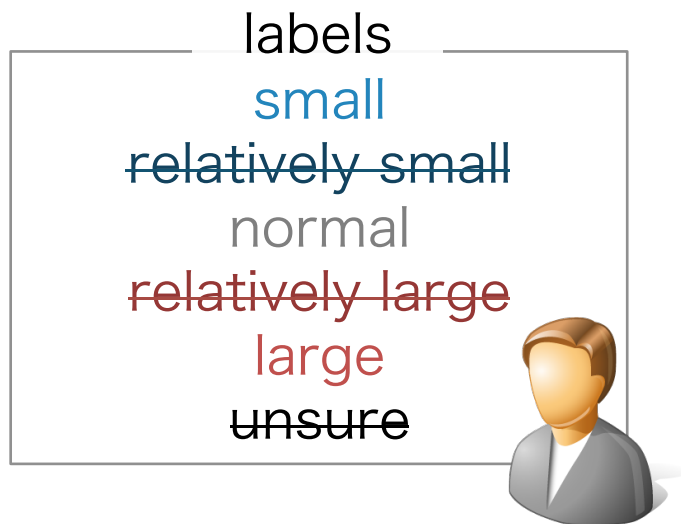
Experimental setup

- We built a gold-standard data set
 - ④ We asked judges to annotate *unsure* when they thought the judgment was highly dependent of the context



Evaluation methods

- gold : 6 labels \Leftrightarrow system : 3 labels
- We remove the sentences with *relatively* and *unsure* label for this evaluation



Results

	Label	P	R	F1	Acc
Clue	large	0.90	0.66	0.78	0.62
	normal	0.59	0.59	0.59	
	small	0.16	0.55	0.36	
Distribution	large	0.86	0.37	0.61	0.59
	normal	0.53	0.91	0.72	
	small	0.22	0.10	0.16	

- The performance is surprisingly good !!
- The clue-based approach was slightly better than the distribution-based

Weak point of distribution-based

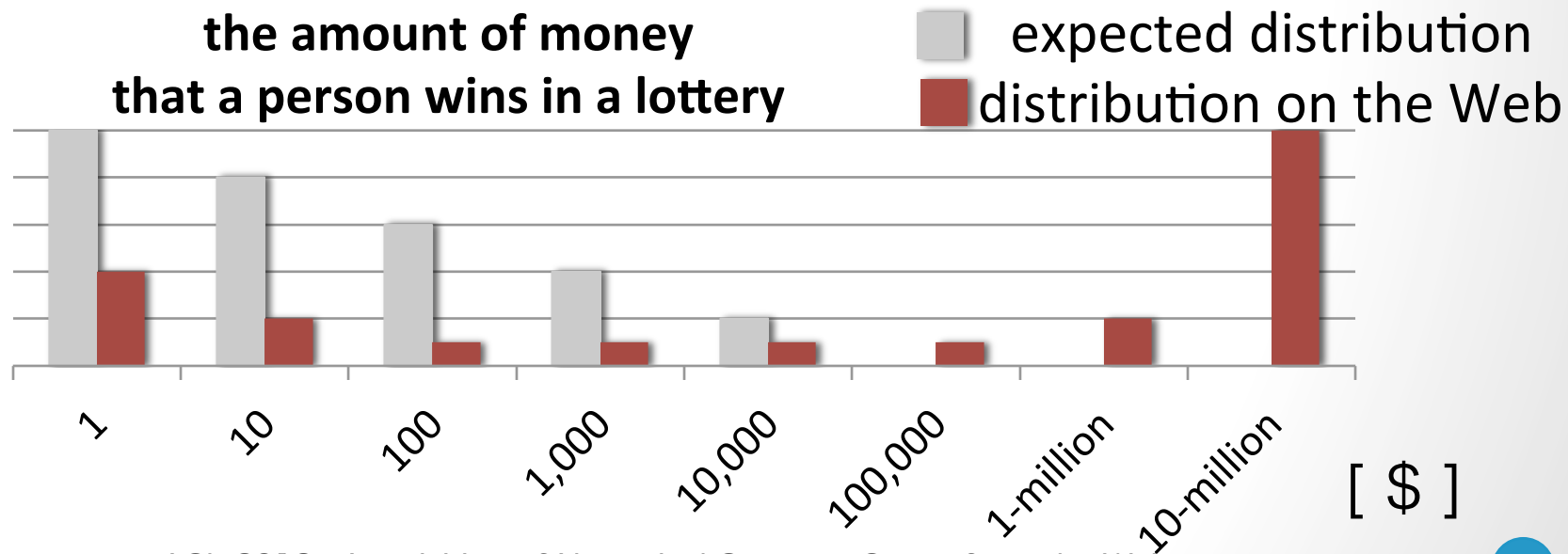
- Sometimes the distribution on the Web is skewed from the 'real' distribution

If I won 10-million-dollar lottery, ...

non fact

Our shop have over 3,000 wonderful items !!

boasting



Output examples (clue-based)

System	Gold	Sentence
small	small	I think that <u>three man</u> can create such a great thing in the world.
large	large	It's <u>above 32 centigrade</u> .

Error analysis (clue-based)

System	Gold	Sentence
small	normal	There are <u>2 reasons</u> .

- Difficulty in judging small.
 - Since some people say, “There are only 2 reasons,” our approach predicted small

Error analysis (clue-based)

System	Gold	Sentence
small	large	<u>Ten or more people</u> came, and my small room was thronged.

- Difficulty in modeling the context
 - 10+ people as the number of guest -> small
 - 10+ people when they are in an small room -> large

Error analysis (clue-based)

System	Gold	Sentence
small	normal	I have <u>two friends</u> who have broken up with their boyfriends recently.

- Difficulty in modeling the context
 - Two as the number of friends -> small
 - Two as the number of friends who have broken up recently -> large

Error analysis (clue-based)

System	Gold	Setence
small	large	The turtle has <u>two heads</u> .

- Lack of knowledge
 - No Web page mentions the number of heads of a turtle
 - It would be better if we could generalize a turtle into an animal

Conclusions

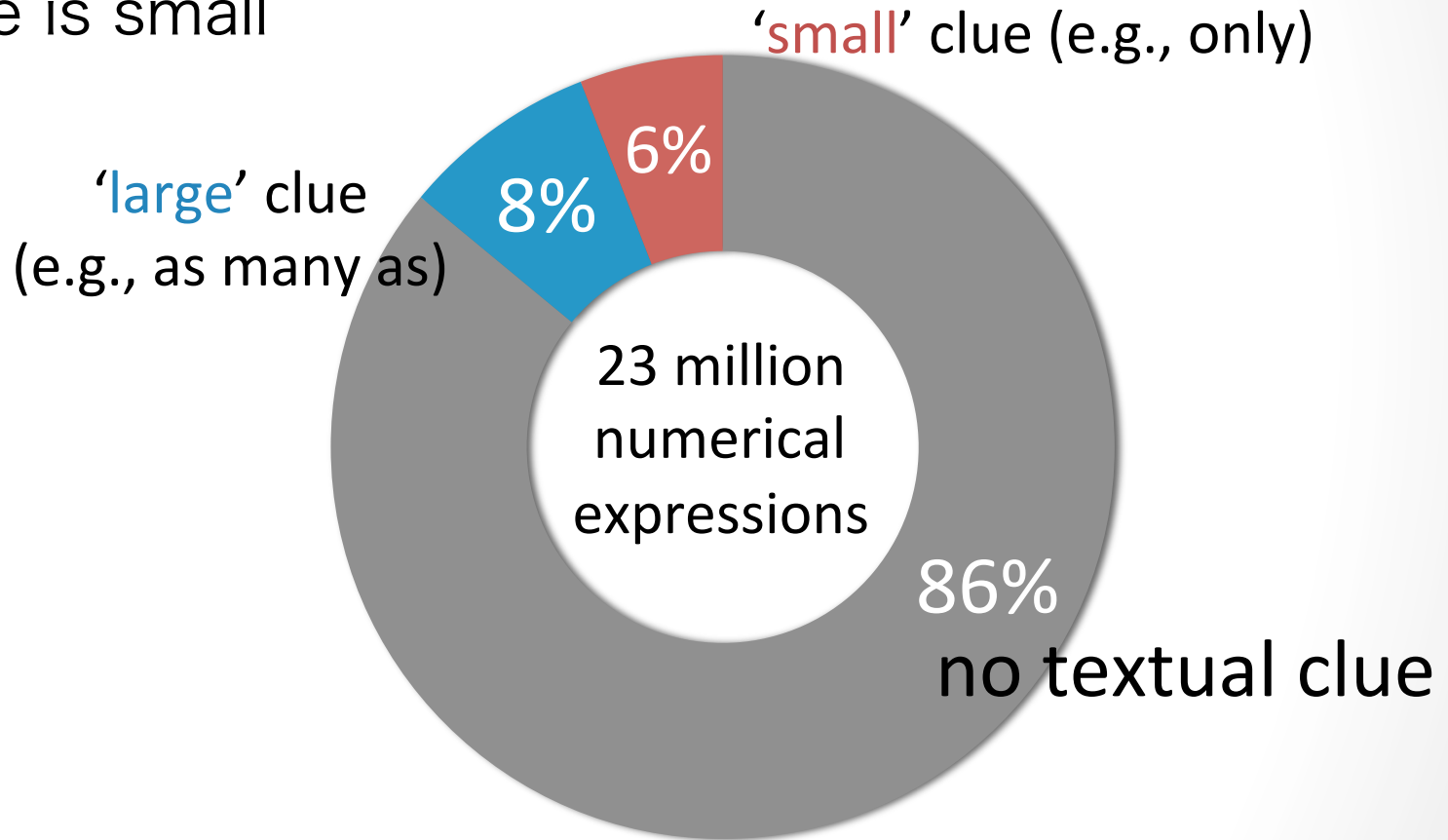
- We proposed novel approaches for acquiring numerical common sense
- The experimental results showed our approaches can successfully judge
- This study is an important step towards a deeper understanding of inferences with numbers !
- Future work
 - Improve the modeling of the contexts
 - Integrate clue-based and distribution-based approaches
 - Solve other types of numerical inferences

Appendix

Related work



Weak point of clue-based

- The number of numerical expressions with textual clue is small



Evaluation methods

- gold : used 6 labels \Leftrightarrow system : used 3 labels
- We employed two criteria for this evaluation

gold 	correct output	
	strict	lenient 
small	small	small
relatively small	do not use this sentence in evaluation	small, normal
normal	normal	normal
relatively large	do not use this sentence in evaluation	normal, large
large	large	large
unsure	do not use this sentence in evaluation	do not use this sentence in evaluation

Results

	Label	P	R	F1	Acc
Clue (strict)	large	0.90	0.66	0.78	0.62
	normal	0.59	0.59	0.59	
	small	0.16	0.55	0.36	
Distribution (strict)	large	0.86	0.37	0.61	0.59
	normal	0.53	0.91	0.72	
	small	0.22	0.10	0.16	
Clue (lenient)	large	0.92	0.78	0.85	0.77
	normal	0.81	0.77	0.79	
	small	0.22	0.70	0.46	
Distribution (lenient)	large	0.89	0.50	0.70	0.76
	normal	0.75	0.94	0.84	
	small	0.27	0.25	0.26	