Semantic annotation in BioCaster

Ai Kawazoe

National Institute of Informatics

Self-introduction

- Name: Ai KAWAZOE（川添愛）
- Country: Nagasaki, Japan
- Affiliation: Project Researcher at NII (2006~)
- Current work: Information Extraction, Ontology design (BioCaster Project, led by Prof. Nigel Collier)
- Doctor in literature (2005, Kyushu University)
- Education: Linguistics (generative grammar, formal semantics)
- Research interest: application of formal studies on language and knowledge to natural language processing

Outline

- Semantic annotation for texts in natural language processing

- Design of semantic annotation

- Issues in BioCaster project --- A case study
  - Designing semantic annotation for disease outbreak information, making use of philosophical/logical foundations

Semantic annotation for texts in natural language processing

What is “Semantic” in “Semantic Annotation”?

- Subfields of linguistics

Syntax

Words: “John”, “got”, “flu”, “a”

Grammatical rules

“John got a flu”   “got flu a John”

(syntactic structure)

[John [got [a flu]]]

Semantics

(semantic representation)

∃x, t [get’(John’, x) & flu’(x) & t<now]

Interpretation

inference

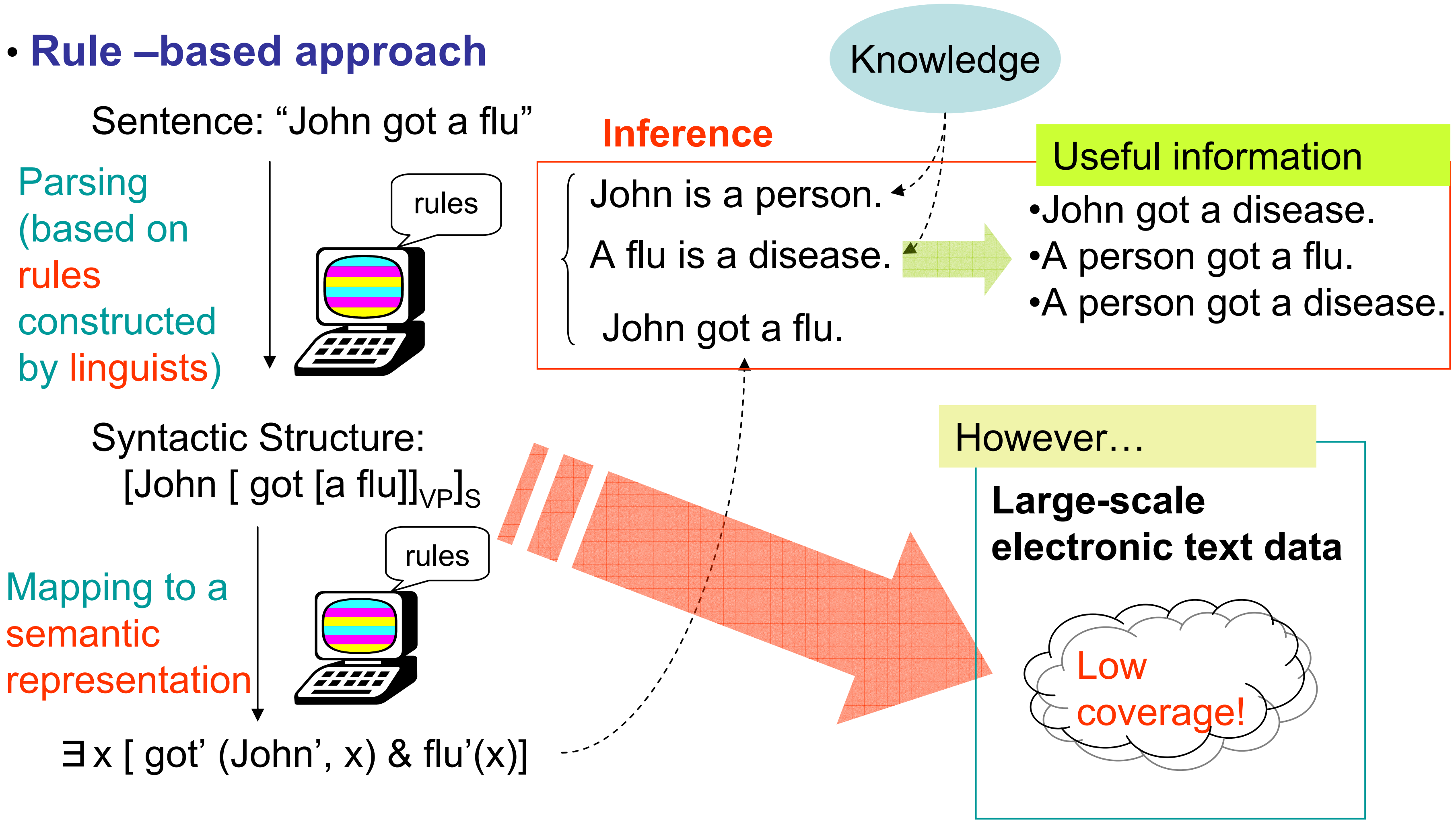
John got a disease.
A person got a flu.
…..

Statistical approach to Natural Language Processing

An example of **Semantic Annotation**

<PERSON>John</PERSON> got a <DISEASE> flu </DISEASE>.

Two approaches in Natural Language Processing (1)



- **Rule –based approach**

Sentence: “John got a flu”

Parsing (based on rules constructed by linguists)

Syntactic Structure:
[John [got [a flu]]$_{VP}$]$_{S}$

Mapping to a semantic representation

$\exists x$ [got’ (John’, x) & flu’(x)]

Knowledge

Inference

John is a person.
A flu is a disease.
John got a flu.

Useful information

- John got a disease.
- A person got a flu.
- A person got a disease.

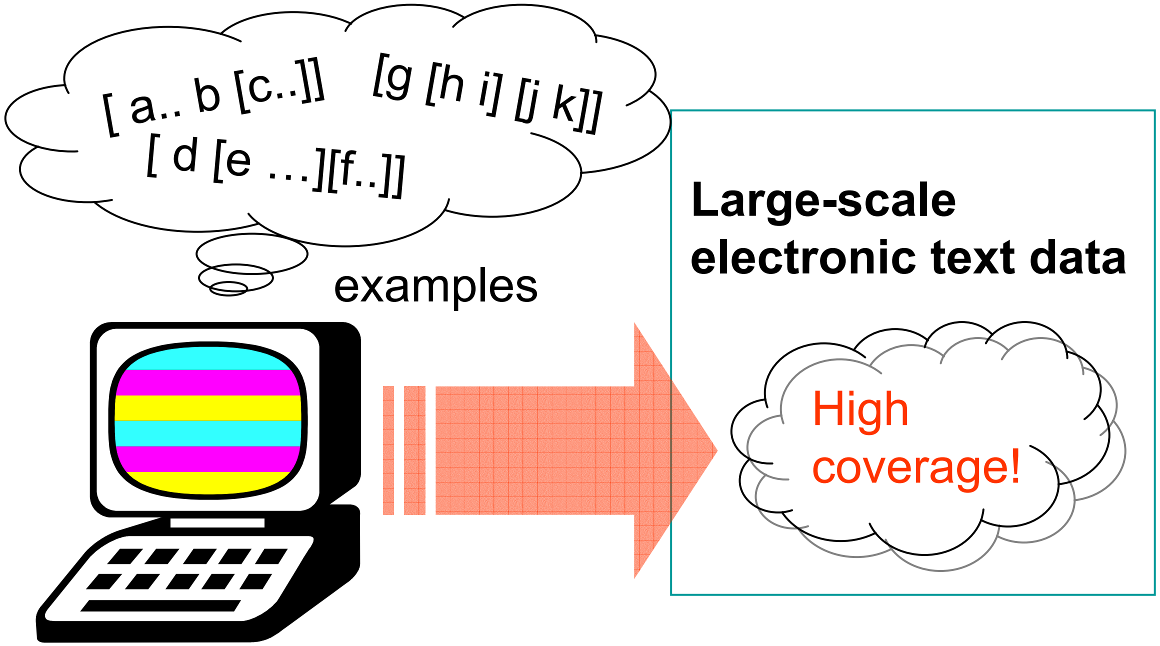
However…

Large-scale electronic text data

Low coverage!

Two approaches in Natural Language Processing (2)

- **Statistical approach (Mid 1990s~)**
- Does not use rules constructed by linguists
- Provides syntactic resources (examples of syntactic structures) to machine
- Grammatical rules are learned by induction from examples



[a.. b [c..]] [g [h i] [j k]]
[d [e …][f..]]

examples

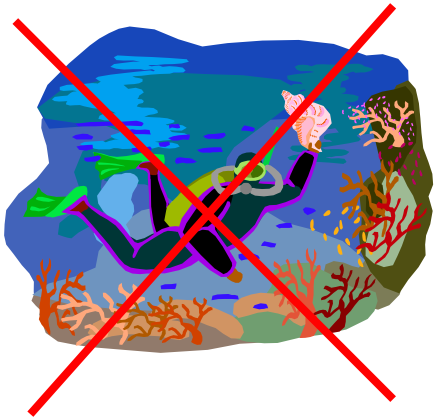
Large-scale electronic text data

High coverage!

However…

- Shallow parsing only, no deep-level semantic representation
- **How can we get useful semantic information?**

Statistical approach and semantic information



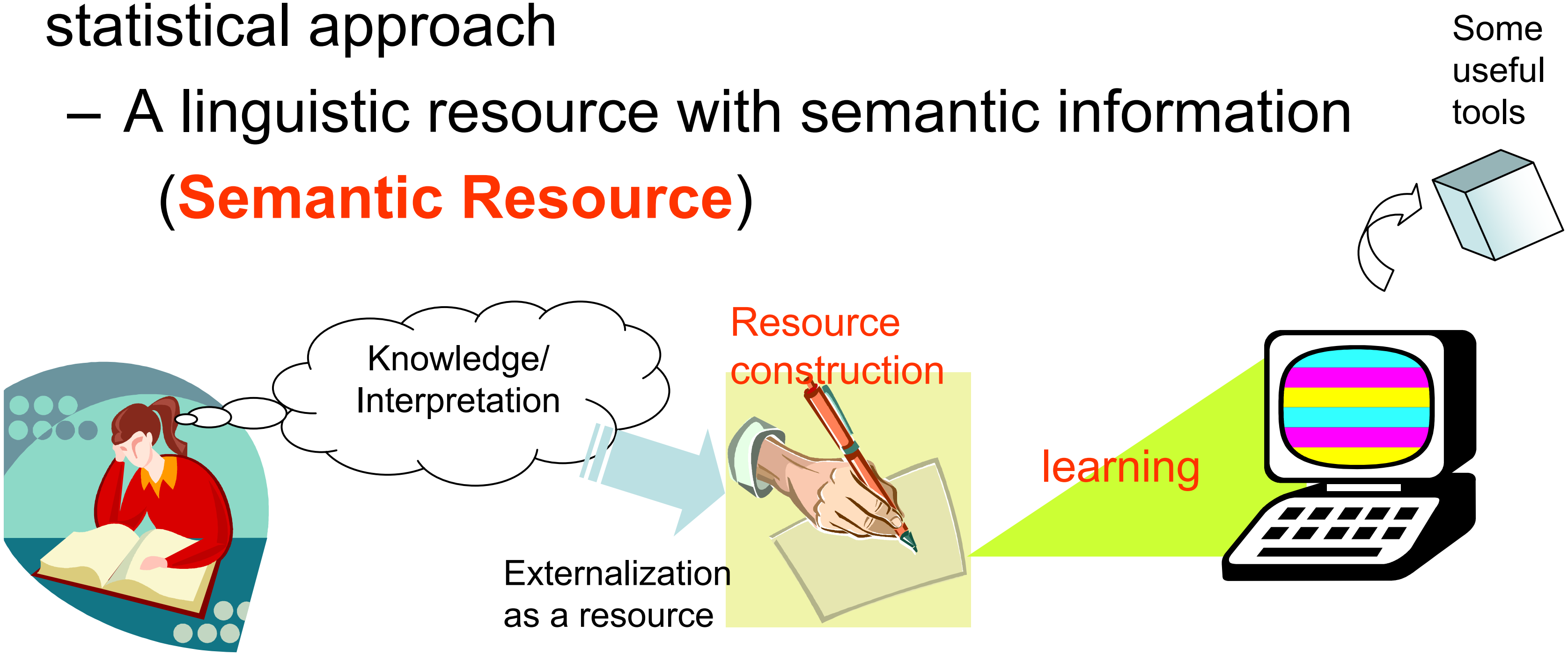
If we cannot obtain a deep syntactic structure----

---then let’s do what we can do in the shallow level !

- Construction of shallow semantic representation
  - Semantic role labeling
  - Named entity recognition
  - Event extraction
  - Ontology induction, etc.

Semantic annotation for constructing “semantic resource”

- One of the important bases for semantic processing in statistical approach
  - A linguistic resource with semantic information (**Semantic Resource**)



A collection of semantic annotation will serve as a semantic resource

Annotation of knowledge & interpretation

- Annotation of real texts with
  1. human’s **knowledge** on the meaning of the text
     - **Annotation for names of person, organization, etc (e.g. MUC-7)**

       <ORGANIZATION>WHO</ORGANIZATION> …

     - **Annotation for technical terms (e.g. GENIA)**

       <PROTEIN>IL-2</PROTEIN>….
       …infected with <VIRUS>H5N1</VIRUS>

  2. human’s **interpretation** of the meaning of the text
     - **Annotation for coreference relations**
     - **Annotation for context-dependent concepts**

       <CASE>A 19-year old girl</CASE> is infected..

       (a case of disease)

Design of semantic annotation

Challenges in designing annotation(1)

- Consistency of annotation is crucial for the performance of the automatic processing of semantic information
- It is not easy to obtain consistency, even with a simple task:

Please annotate names of people !

Do I have to annotate "Charles de Gaulle" in "the Charles de Gaulle airport" ???

Confusion

Sir [Arthur Conan Doyle]
Or
[Sir Arthur Conan Doyle] ???

Personal rules?

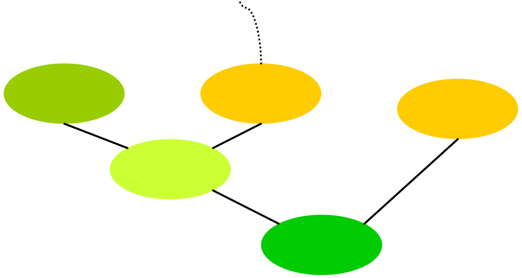
Annotator

Tools for semantic annotation

- Annotation can contain rich information ⟹ complexity
- Annotation is a kind of language ⟹ ambiguity

… <annotation> TEXT </annotation> …

How to regulate?



Ontology provides definitions of markable concepts

Annotation guidelines defines a schema for annotation

Ontology
+
Annotation guidelines

These should be well-designed at the early stage!

What is necessary to design good annotation schema

Designer

1.

Clarification of

•Definitions of ‘Markable’ concepts

•User needs

2.

Clear and intelligible presentation of annotation schema for annotators

Annotator

Difficult to meet in an ad-hoc way… we want a principled way if available

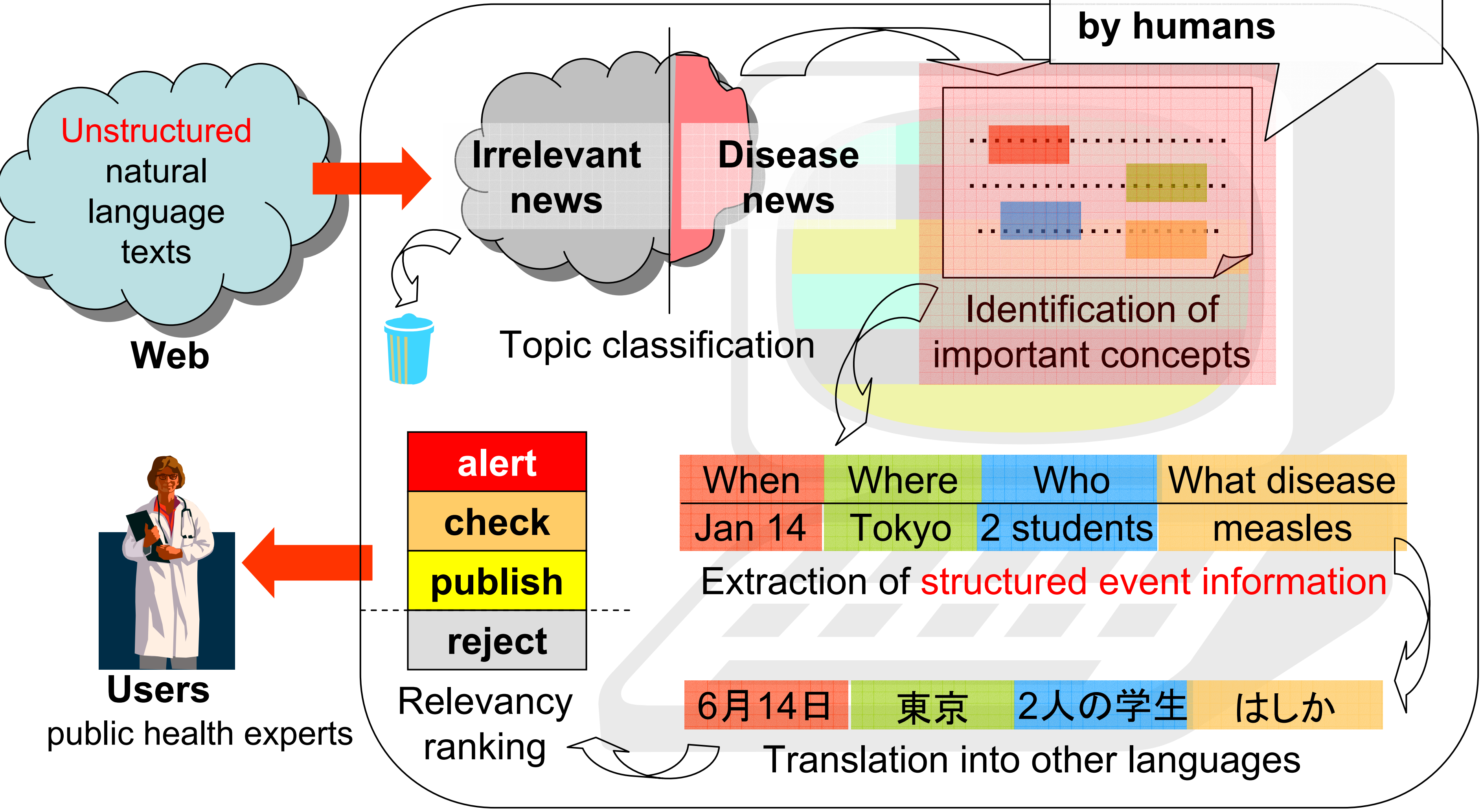
Our claim:

Philosophical / Logical / Linguistic considerations are useful tools to design annotation schema

A Case Study: Semantic Annotation in BioCaster Project

BioCaster system: Overview

- An intelligent system to monitor disease outbreak information on the Web



Machine learning from semantic annotation created by humans

Unstructured natural language texts

Web

Irrelevant news

Disease news

Topic classification

Identification of important concepts

When	Where	Who	What disease
Jan 14	Tokyo	2 students	measles

Extraction of structured event information

6月14日	東京	2人の学生	はしか

Translation into other languages

alert

check

publish

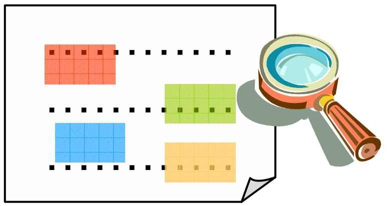
reject

Relevancy ranking

Users

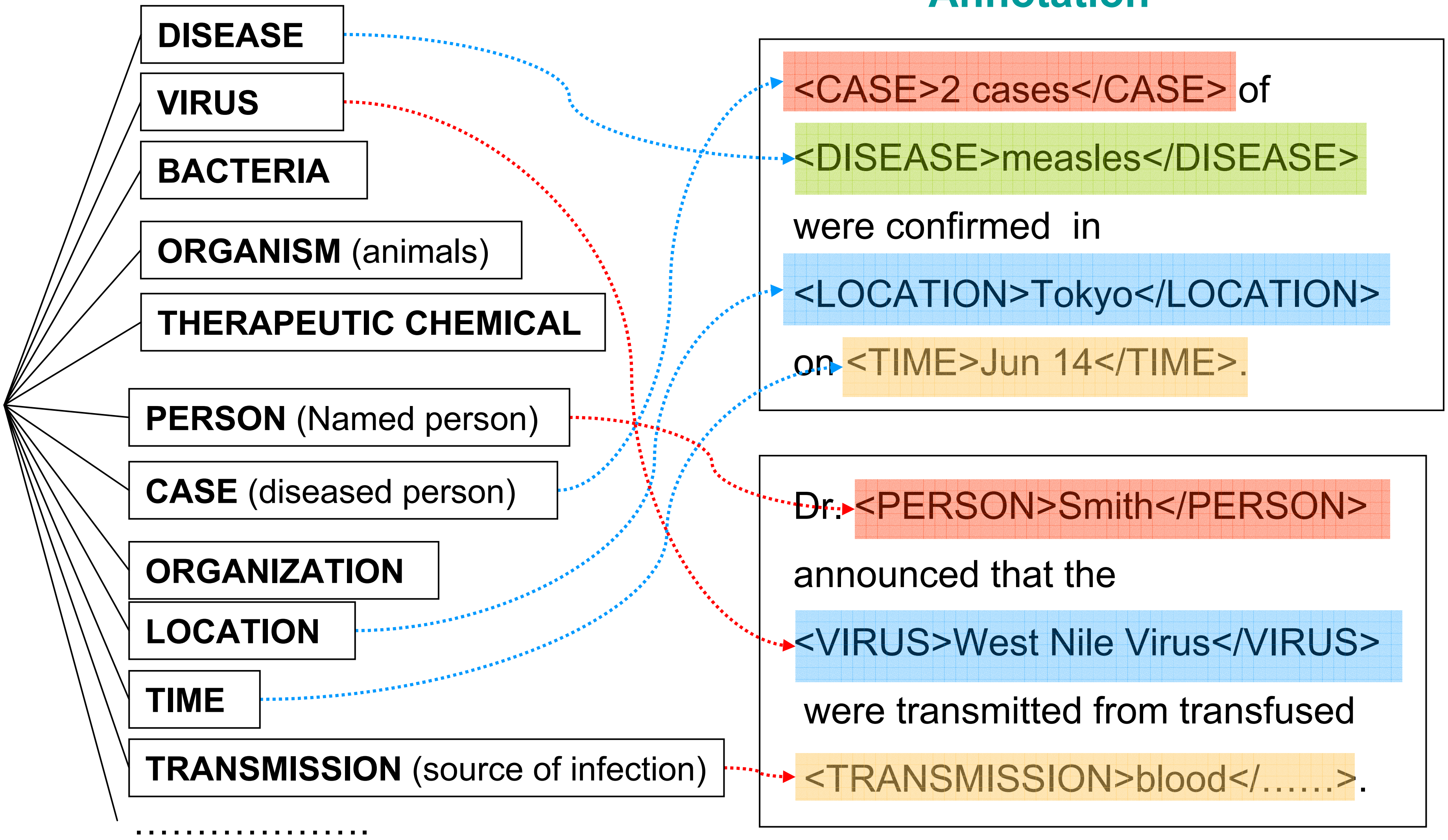
public health experts

‘Markable’ concepts (1st)



Ontology

Annotation



DISEASE

VIRUS

BACTERIA

ORGANISM (animals)

THERAPEUTIC CHEMICAL

PERSON (Named person)

CASE (diseased person)

ORGANIZATION

LOCATION

TIME

TRANSMISSION (source of infection)

…………………

<CASE>2 cases</CASE> of <DISEASE>measles</DISEASE> were confirmed in <LOCATION>Tokyo</LOCATION> on <TIME>Jun 14</TIME>.

Dr. <PERSON>Smith</PERSON> announced that the <VIRUS>West Nile Virus</VIRUS> were transmitted from transfused <TRANSMISSION>blood</……>.

Problems in 1st annotation experiment (1)

A WHO laboratory confirmed that Mr.**Yamada** was infected with the virus

I think it is a **CASE** (diseased person) since “Mr. Yamada” here is sick

I think it is a **PERSON** (named person) since “Mr. Yamada” here is mentioned by name

Annotator 1

Inconsistent!

Annotator 2

Problems in 1st annotation experiment (2)

Victims contract the virus from close contact with infected **birds**

I think it is a **TRANSMISSION** (source of infection) since it transmitted virus to others

I think it is an **ORGANISM** since it is a mention to animals

Annotator 1

Inconsistent!

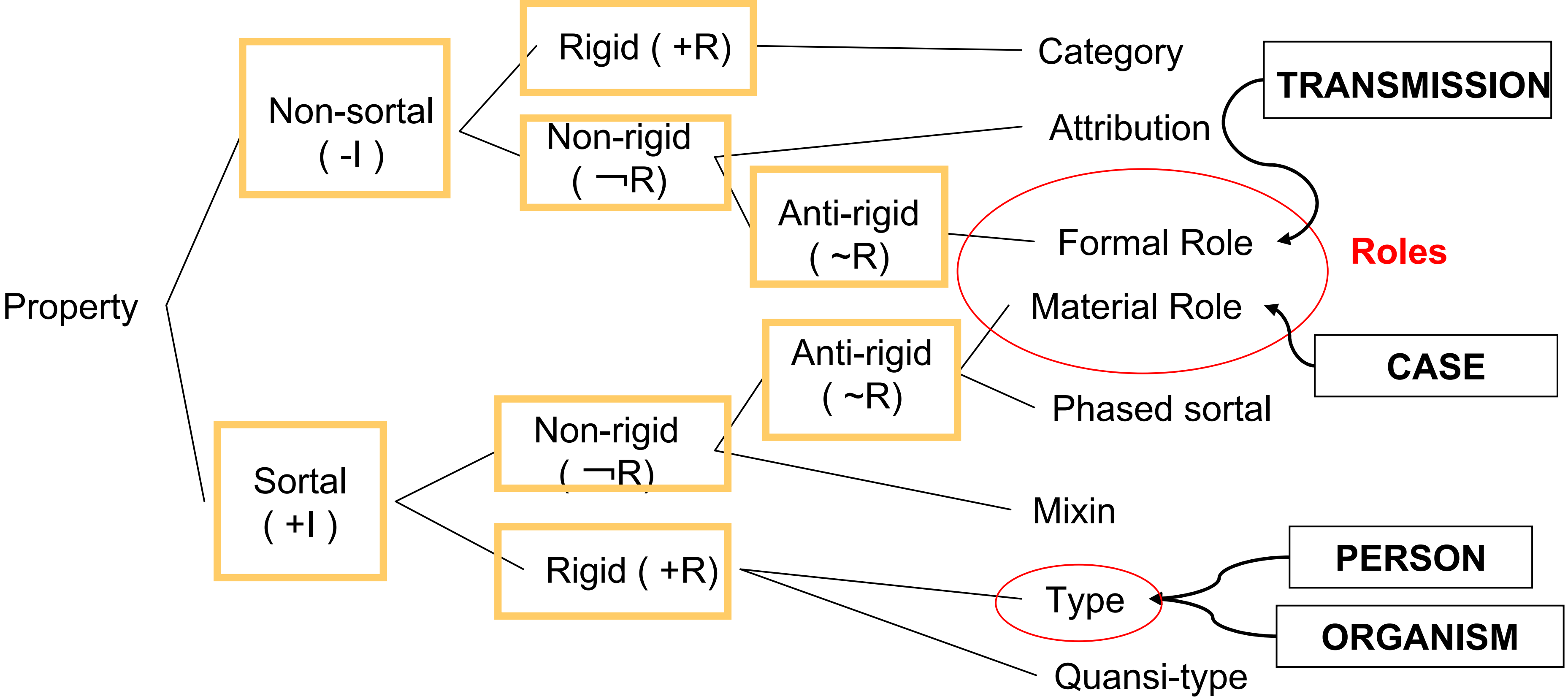
Annotator 2

Reanalysis of “markable” concepts (1)

- **Method:**

Classification of concepts by Guarino and Welty (2000a, b)

Based on fundamental philosophical notions



Property

Non-sortal (-I)

Rigid (+R)

Category

Non-rigid (¬R)

Attribution

Anti-rigid (~R)

Formal Role

Material Role

Roles

TRANSMISSION

CASE

Sortal (+I)

Non-rigid (¬R)

Anti-rigid (~R)

Phased sortal

Mixin

Rigid (+R)

Type

Quansi-type

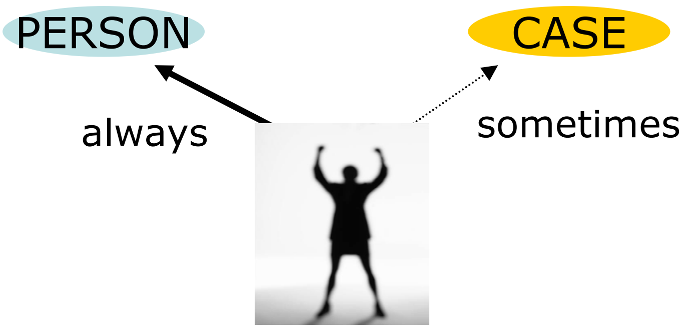
PERSON

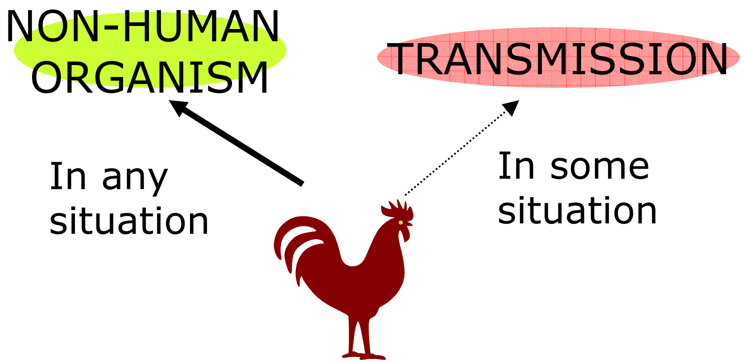
ORGANISM

Reanalysis of “markable” concepts (2)

Now we know ---

- Role concepts are the problematic ones!
- Role concepts are basically ambiguous --- something which has a role belongs to some Type concept.

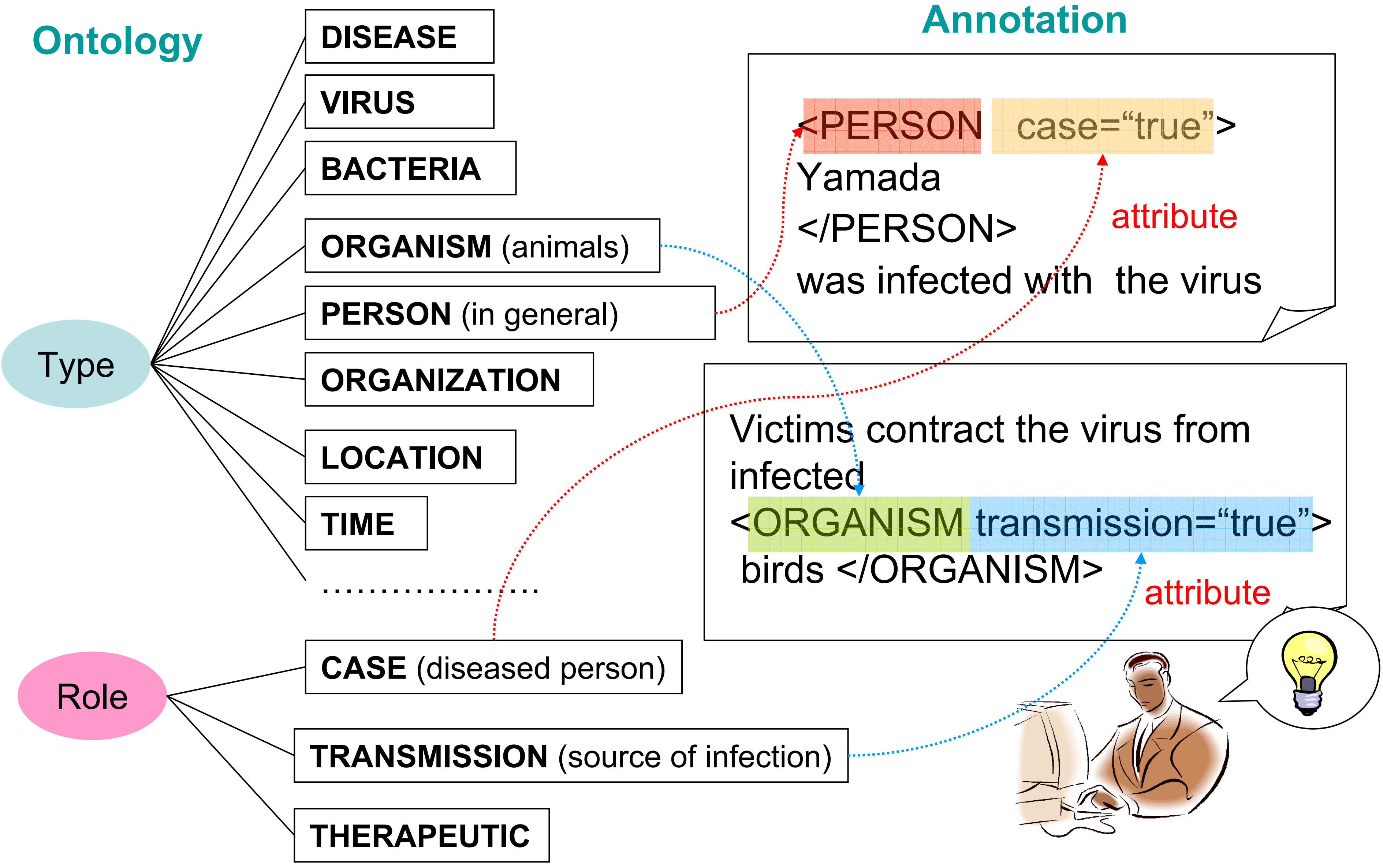




- We should make a clear distinction between Roles and Types in the ontology and the annotation schema!

- “Therapeutic chemical” is also identified as a role --- we can prevent problems in advance.

Change of the annotation schema



Ontology

Type

DISEASE

VIRUS

BACTERIA

ORGANISM (animals)

PERSON (in general)

ORGANIZATION

LOCATION

TIME

…………………

Role

CASE (diseased person)

TRANSMISSION (source of infection)

THERAPEUTIC

Annotation

<PERSON case=“true”>
Yamada
</PERSON>
was infected with the virus

attribute

Victims contract the virus from infected
<ORGANISM transmission=“true”>
birds </ORGANISM>

attribute

Results of automatic entity recognition (1st corpus vs. 2nd corpus)

	1st (F-score)	2nd (F-score)
Overall	76.96	79.96 (+3)
PERSON	54.95	65.63 (+11.33)
PERSON (case="true")	53.17 (CASE)	66.28 (+12.46)
ORGANISMS	68.0	73.21 (+5.21)

Our other works with similar approach

- Annotation of epistemology-loaded expressions (e.g.“suspected case”)
- Coreference annotation
- Problems of polysemy

Application of

•**Philosophical**
•**Logical**
•**Linguistic**

foundations

Combination of

Highly-abstract formal studies on knowledge

+

Knowledge Engineering

Conclusion

- Semantic annotation is a technology to construct a semantic resource for machine understanding of “meaning” of natural language

- A case study in BioCaster project --- Philosophical/logical methodology is useful in designing annotation schema

- Future issues --- Integration of “principled” ways to design good annotation schema, by applying foundations of abstract, formal studies on knowledge and language.

Thank you for your attention!