KR-MED 2006 "Biomedical Ontology in Action"
November 8, 2006, Baltimore, Maryland, USA

# The development of a schema for the annotation of terms in the BioCaster disease detecting/tracking system

**Ai Kawazoe[*1], Ph.D., Lihua Jin[*1], Ph.D., Mika Shigematsu[*3], M.D.,**
**Roberto Barrero[*2], Ph.D., Kiyosu Taniguchi[*3], M.D., Nigel Collier[*1], Ph.D.**
**[*1]National Institute of Informatics, Hitotsubashi 2-1-2 Chiyoda-ku Tokyo, JAPAN**
**[*2]National Institute of Genetics, Yata 1111 Mishima Shizuoka, JAPAN**
**[*3]National Institute of Infectious Diseases, Toyama 1-23-1 Shinjuku-ku Tokyo, JAPAN**
[*1]{zoeai,lihua-jin,collier}@nii.ac.jp, [*2]rbarrero@genes.nig.ac.jp,
[*3]{mikas,tanigk}@nih.go.jp

*Amid growing public concern about the spread of infectious diseases such as avian influenza and SARS, there is an increasing need for collecting timely and reliable information about disease outbreaks from natural language data such as online news articles. In this paper we introduce BioCaster, a text mining-based system for infectious disease detection and tracking currently being developed, and discuss the development of a domain ontology and schema for the annotation of terms. In particular we focus on the comparison between two approaches, 1) a traditional task-oriented approach with a simple schema that does not strictly follow ontological principles, and 2) a formal approach which is ontologically well-founded but adds extra requirements to the annotation schema. We report on several critical problems that were highlighted by an entity annotation experiment, attributable to the purely task-oriented ontology design. A second experiment based on a formally constructed ontology produced improved annotation results despite the apparent complexity of the annotation schema.*

## 1. INTRODUCTION

As shown by the recent outbreak of Severe Acute Respiratory Syndrome (SARS) and emerging cases of avian influenza, infectious diseases have the potential to spread rapidly through person-to-person transmission within densely populated areas and across country borders through international air travel. The first line of defense against rapidly spreading diseases is surveillance, led by the World Health Organization (WHO) and national health authorities. Catching an outbreak earlier has clear implications for both morbidity and mortality as well as the feasibility of containment [1]. However a lack of surveillance system infrastructure in Southeast Asia, which is currently the focus of an avian H5N1 epidemic is seen as hindering control efforts. In addition to traditional surrogate methods such as reporting notifiable diseases and over-the-counter (OTC) sales monitoring, public health experts are increasingly considering news and other reports available on the World Wide Web (Web) as a cost-effective means of helping to find and track early cluster cases, enabling a timely and appropriate response. Such *rumour-based* information may be of particular value for assessing possible outbreaks in areas where formal reporting procedures are absent or not well established.

Several major challenges exist in locating Web-based information in a timely manner using traditional search methods: (1) the massively increasing volume of dynamically changing unstructured news data available on the Web makes it extremely difficult to obtain a clear picture of an outbreak in a timely manner, (2) the large-scale republication of reports from centralized news agencies requires redundancy to be identified and removed, (3) the initial reports of an outbreak are contained in only a few news articles which will usually be overlooked by traditional search engines which use keyword indexing, (4) the first reports of an infectious disease will often be reported in local news media which are only available in the local language. Experience has shown that this requires computer systems to have at least a partial understanding of the domain through ontologies, term lists and databases as well as specialized multilingual resources.

To address the information needs in the domain of infectious disease outbreaks, standard Information Extraction technology has been adapted for retrospective archive search [2] but only a few systems are currently actively deployed with the most prominent being the Global Public Health Intelligence Network (GPHIN) [3], a successful but semi-closed system used by the WHO. We are now developing BioCaster, a text mining system based on an openly available multilingual ontology for proactive notification about priority disease outbreaks. A key component of the BioCaster system is the use of automated learning methods to identify novel entities and events using features derived from annotated examples in a multilingual collection of news articles. The initial target languages are English, Japanese, Vietnamese and Thai.

In our early development of BioCaster it became clear that we needed a rigorous schema for markable entities. Since the system relies on high quality human annotated training data for constructing

named entity recognizers (NERs), any inconsistency introduced into the annotation schema by ontological inconsistencies should be harmful for annotation performance, both human and machine. Surprisingly while there have been several studies on the mapping problem between terms and coding systems such as the UMLS Metathesaurus [4] as well as biomedical annotation experiments [5] [6] [7] there have been to the best of our knowledge no studies conducted into the method by which new domain models suitable for biomedical text mining should be organized. We report here on our initial experience which showed that the task-oriented annotation schema based on a poorly-considered domain ontology can indeed be harmful to accuracy. Re-organizing this schema using well founded ontological principles produced better results, despite the added complexity.

## 2. USER NEEDS

Epidemiologists are concerned with the circumstances in which diseases occur in a population and the factors that influence their incidence, spread, recognition and control. Our initial discussions with domain experts at the National Institute of Infectious Diseases revealed several common scenarios for gathering information from Web news including cases involving the spread of a communicable disease across international borders and the contamination of blood products. From these initial discussions we collected examples of early outbreak news reports and compiled a list of significant entity classes which included DISEASE[1], CASE, LOCATION SYMPTOM, TIME, DRUG, etc.

Subsequent follow up discussions and examination of the literature revealed that we can categorize these concepts according to the information needs of the scientists as shown in Table 1.

Genetic epidemiology adds another dimension to the information needs as the genetic makeup of the host plays a key role in determining susceptibility or resistance to pathogens. We therefore chose to add in a further level of detail about the host which includes genes and their products, identified with a §. Finally we had 19 categories of concepts which we want to identify in news texts (Table 2).

## 3. CONSIDERATION ON TWO APPROACHES

At this stage we were aware that some of the important concepts in Table 2 are contextually-dependent and intrinsically different from others. For example, CASE and TRANSMISSION represent roles (discussed in [8] [9] [10] [11] among others) which are dependent on the existence of events they participate in, while most others, such as PERSON, BACTERIA, and NON_HUMAN, represent types.

We had two options for constructing the ontology and annotation schema, according to how to deal with concepts of a different nature. The first approach is rather task-oriented. Here we do not make any distinction between context-dependent concepts and others. This results in a somewhat simpler ontology: all categories of concepts are represented as classes which follow a disjoint entity class principal that has been the underlying premise of NERs. The corresponding annotation schema will also be simpler, since instances of context-dependent classes are annotated in the same way as those of other classes, e.g.

<NAME cl="PERSON">Kofi Annan</NAME>
<NAME cl="CASE">a 12 year-old girl</NAME> infected with H5N1

(The details of this schema will be given in the next section.) In this task-oriented approach, we can annotate exactly what the event frame needs to identify. For example, we can exclude from annotation non-named, non-case mentions, which we are not interested in. A defect of this approach is that it is not ontologically well-founded.

The alternative approach is a more formal one where we make a clear distinction between context-dependent concepts and others, based on well-founded ontological principles. The result is likely to be a more complex ontology in which context-dependent concepts have a different status from other concepts. The corresponding annotation schema will also be more complex as well, since roles are annotated in a different way from those of entity classes. In order to achieve ontological consistency we also need to annotate more mentions than the former approach, including those that will not instantiate event frames.

From the two approaches above, out of expediency we chose the former for the first annotation experiment. The reason being that it seemed easier for annotators and that we could find almost no precedent works in named entity annotation which dealt with formal analysis of entities and role concepts.

## 4. ANNOTATION EXPERIMENT 1

### 4.1 Method

Based on the list of categories of concepts in Table 2, we constructed the ontology shown in Figure 1. Note that CASE and TRANSMISSION, which represent

[1] We will adopt here the notation of using all upper case for domain entity classes.

| Focus | Description | Example properties | Concept types |
|---|---|---|---|
| Agent | Pathogens | Infectivity, pathogenicity, virulence, incubation period, communicability | VIRUS, BACTERIA, PARASITE*, FUNGI* |
| Transmission | The delivery or dispersal method | Dermal, oral, respiratory | TRANSMISSION |
| Host | Persons carrying a disease | Age, gender, occupation, | CASE, SYMPTOM, DISEASE, ANATOMY, DNA§, RNA§, PROTEIN§ |
| Environment | Location and climate | Large population centre, enclosed building, mass transport system, rural village | LOCATION, TIME |
| * Not included in the current schema | | | |
| § Genetic level entities | | | |

*Table 1 Categorization of concepts*

| Classes | Examples | Description |
|---|---|---|
| ANATOMY | *liver, pancreas, nervous system, eLa cel,* | Body parts including tissues and cells |
| BACTERIA | *Escherichia coli O157*, *tubercle bacillus* | Eubacteria |
| CASE | *a 35-year-old woman, the third case* | Confirmed cases of diseases |
| NT_CHEMICAL | *beryllium, organophosphate pesticide* | Chemicals intended for non-therapeutic purposes *[1] |
| T_CHEMICAL | *Relenza, immunosuppressive drug, oseltamivir* | Chemicals intended for the treatment of diseases*[1] |
| CONTROL | *stamping out, screening, vaccination* | Control measures to lower the risk of transmission of a disease |
| DISEASE | *H5N1 avian influenza, SARS, cholera* | A deviation in the normal functioning of the host caused by a persistent agent (pathogen) or some environmental factor |
| DNA | *Sp1 site, triple-A, c-jun gene* | Includes the names of DNAs, groups, families, molecules, domains and regions*[2] |
| LOCATION | *Viet Nam, Jakarta, Sumatra Island, Asia* | A politically or geographically defined location*[3] |
| NON_HUMAN | *civet cats, poultry, flies* | Multi-cell organism other than humans, i.e. "animals" |
| ORGANIZATION | *the Ministry of Health, WHO, Pasteur Institute* | Corporate, governmental, or other organizational entity*[3] |
| PERSON | *Jean Chretien, Murray McQuigge* | A named person or family |
| PRODUCT | *botulism antitoxin, Influenza vaccine* | Biological product, (e.g. vaccines, immune sera) |
| PROTEIN | *STAT, RNA polymerase II alpha subunit* | Includes the names of proteins, groups, families, molecules, complexes and substructures*[2] |
| RNA | *IL-2R alpha transcripts, TNF mRNA* | Includes the names of RNAs, groups, families, molecules, domains and regions*[2] |
| SYMPTOM | *cough, fever, dehydration, convulsion* | Alterations in the appearance of a case due to a disease |
| TIME | *Tue Jan 3, winter, March, since October, 2003* | Temporal expressions that can be anchored on a timeline*[4] |
| TRANSMISSION | *HIV-tainted blood products, BSE-infected cows* | Source of infection |
| VIRUS | *Ebola virus, HIV* | Viruses such as HIV, HTLV, EBV *[2] |
| Descriptions marked with *1 , *2, *3, *4 are based on those in MeSH [12], GENIA ontology [13], MUC-7 [14], and HUB-4 [15], respectively. | | |

*Table 2 List of classes of markable concepts*

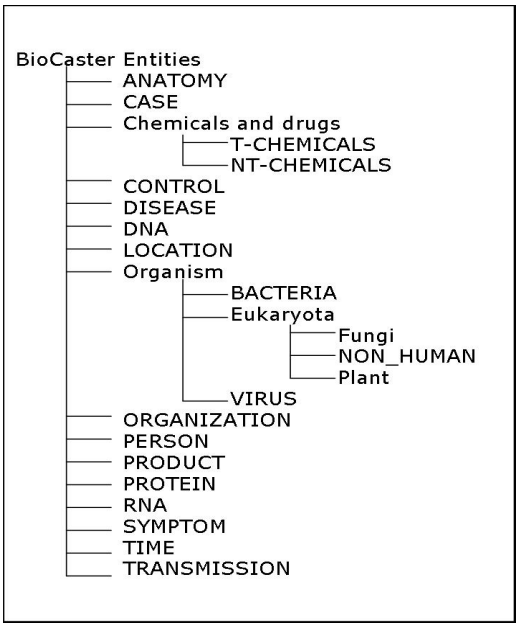


*Figure 1 Initial domain ontology (simplified)*

roles, have the same status as other classes since we adopted the task-oriented approach as discussed in the last section. We developed annotation guidelines to annotate non-overlapping mentions related to the classes in news articles and hired two PhD informatics students as annotators. After 1-week of training consisting of guideline review, case study discussions and test cases, we started the annotation process with 200 news articles taken from domain sources, including WHO epidemic reports, IRIN, and Reuter news.

In order to restrict the markable mentions to exactly those that we aimed to identify with the text mining system, we defined CASE as the class of confirmed cases which are unnamed, and PERSON as the class of named persons who are not cases. We considered this would narrow down the number of markable mentions since unnamed mentions for non-cases need not be annotated. We also instructed annotators to markup only the single most appropriate class, prohibited multiple classes. An example of annotated text is shown below:

> The <NAME cl="ORGANIZATION">Ministry of Health</NAME> in <NAME cl="LOCATION"> Indonesia</NAME> has today confirmed <NAME cl="CASE">a fatal human case</NAME> of <NAME cl="DISEASE">H5N1 avian influenza</NAME>. <NAME cl="CASE">A 27-year-old woman</NAME> from <NAME cl="LOCATION">Jakarta</NAME> developed symptoms on <NAME cl="TIME">17 September</NAME>. She contracted the virus from close contact with infected <NAME cl="TRANSMISSION">birds</NAME>.

In the annotation schema used in the example above, the attribute *cl* takes the entity class label as its value. For example "<NAME cl="PERSON">Kofi Annan</NAME>" means that the entity mentioned by "Kofi Annan" is *related* to the class PERSON. The reason for using this rather vague expression is to cover two relations between mentioned entities and the ontology we want to describe. The first is "is an instance of", and the other one is "is a subclass of". Some of the markable texts mention a particular and others mention a universal. For example, names of persons, locations and organizations are usually used to refer to a particular, whereas names of chemical substance, viruses and proteins are often used to refer to universals. This is one of the factors which makes ontology-based annotation a complicated process. It should be noted though that we intend to work towards a clear distinction between the two relations in future work.

### 4.2 Annotation results and problems

During the first annotation experiment, we had many problem reports form annotators, and found a significant number of inconsistencies in the annotation results. Most of the problems could be traced back to poor design of the domain ontology and the annotation schema. Follow up analysis on the corpus yielded the following symptoms of error:

- Gaps in the annotation schema shown by the existence of mentions to entities which it is desirable to annotate but the annotation schema does not cover.
- Ambiguity between context-dependent concepts and context-independent ones
- Idiosyncratic annotations which are forced on annotators due to the disjoint entity class principal.

**Gaps in the annotation schema**

At the initial stage of our analysis we considered that distinguishing CASE (as confirmed cases of a disease which are unnamed humans) from PERSON (named persons who are not cases of a disease) was rather natural, since CASE entities are in general anonymous. However, in the news articles there were some examples where cases were mentioned by name as follows:

E1 Tests carried out in a UK laboratory confirmed that **M.A** and **F** died from the H5N1 strain[2]

In addition, we found that there were more frequent mentions of putative cases than we had expected.

[2] In this example we only show initials of the victims' names.

These mentions were often annotated as CASE by annotators although we restricted the scope of this class only to confirmed cases.

E2 **a Taiwanese** is suspected to have died of SARS

Follow up discussions with public health experts revealed that mentions of putative cases are important, especially in the early stages of disease outbreaks, and we concluded that they should be identified by the system. However, the existing framework made them difficult to capture.

**Ambiguity caused by context-dependent concepts**
One of the classes which confused annotators most was TRANSMISSION (source of infection). Below are typical examples of problematic cases.

E3 Victims contract the virus from close contact with infected **birds**
E4 There is no known cure for Ebola, which is transmitted via infected **body fluids**
E5 An Irish woman infected with Hepatitis C by a contaminated **blood product**
E6 18 hospitalized after consuming **chapattis**

Annotators had a problem in annotating 'birds' in E3 since those can be classified as both TRANSMISSION and NON_HUMAN (animals). 'Body fluid' in E4 is also ambiguous between TRANSMISSION and ANATOMY (body parts), and also 'blood product' in E5 is ambiguous between TRANSMISSION and PRODUCT (biological product). Most of the TRANSMISSION instances found in the text were those which could be categorized as NON_HUMAN, and the cases which belonged only to TRANSMISSION, such as 'chapattis' in E6, were very few.

**Idiosyncratic annotations due to the disjoint entity class principal**

E7 <NAME cl="PERSON">**Hudd**</NAME> has written several books on music hall and Variety...
E8 Doctors later diagnosed <NAME cl="CASE">**Hudd**</NAME> with a chest infection...

In the example above, it is clearly undesirable that the same entity is related to PERSON in E7 and CASE in E8. Although the annotator was aware of the choices the principal of disjoint classes forced a choice.

### 4.3 Empirical results from training an NER

We trained a support vector machine [13] (for details, see Takeuchi and Collier [14]) for named entity recognition based on the annotated corpus of 200 news articles. 10-fold cross validation experiments were performed using TinySVM[3]. A -2/+1 features window was used that included surface word, orthography, biomedical prefixes/suffixes, lemma, head noun and previous class predications. The F-score for the all classes in Table 2 was 76.96. Among the problematic classes were found to be PERSON, CASE and NON_HUMAN (many instances of which had ambiguity with TRANSMISSION) which had F-scores below our expectation: PERSON (54.95), CASE (53.17), NON_HUMAN (68.0).

## 5. ANNOTATION EXPERIMENT 2

### 5.1 Re-examination of the approach

Although we chose the task-oriented approach for its simplicity and ease of implementation the results from automatic NER and subsequent corpus analysis revealed that problems arose because we made no clear distinction between context-dependent and context-independent classes. We decided to take an alternative, formal and linguistically-sound approach, and distinguish context-dependent concepts from others in both the ontology and the annotation schema.

### 5.2 Classification of concepts

The first step was to use the classification method proposed by Guarino and Welty ([9] and [10]) which is based on meta-properties (rigidity, identity, dependency), in order to classify categories of concepts in Table 2. Definitions of the meta-properties we used are as follows:

**<Rigidity>** ([10], p.4)
**rigid property $\phi$ (+R)**: $\forall x\ \phi(x) \rightarrow \Box\ \phi(x)$
**anti-rigid property $\phi$ (~R)**: $\forall x\ \phi(x) \rightarrow \neg\Box\ \phi(x)$

**<Identity>** ([10], p.5)
**Identity Condition (IC)**: An identity condition is a formula $\Gamma$ that satisfies either of the followings[4]:

[3] Available from http://cl.aist-nara.ac.jp/~taku-ku/software/TinySVM
[4] In [9], further restrictions are added in order to avoid 1) the case where the necessary IC definition becomes trivially true regardless of the truth value of the formula x=y and 2) the case where $\Gamma(x, y, t, t')$ is false and that makes the sufficient IC definition trivially true.

| | rigidity | identity (supplying) | identity (carrying) | dependency | classification |
|---|---|---|---|---|---|
| ANATOMY | +R | +O | + I | - D | Type |
| BACTERIA | +R | +O | + I | - D | Type |
| CASE | ~R | - O | + I | +D | Material Role |
| NT_CHEMICAL | ~R | - O | + I | +D | Material Role |
| T_CHEMICAL | ~R | - O | + I | +D | Material Role |
| CONTROL | ~R *1 | - O*2 | + I | +D | Material Role |
| DISEASE | +R | +O*3 | + I | +D | Type |
| DNA | +R | +O | + I | - D | Type |
| LOCATION | +R | +O | + I | - D | Type |
| NON_HUMAN | +R | +O | + I | - D | Type |
| ORGANIZATION | +R | +O | + I | - D | Type |
| PERSON | +R | +O | + I | - D | Type |
| PRODUCT | +R | +O | + I | +D | Type |
| PROTEIN | +R | +O | + I | - D | Type |
| RNA | +R | +O | + I | - D | Type |
| SYMPTOM | +R | +O | + I | +D | Type |
| TIME | +R | +O | + I | - D | Type |
| VIRUS | +R | +O | + I | - D | Type |
| TRANSMISSION | ~R | - O | - I | +D | Formal Role |

*1 We consider that this class is anti-rigid, since it is possible that an action which is an instance of CONTROL in the current world is not an instance of CONTROL in some other accessible world. The same action may be conducted for different purposes in different worlds.

*2 This class includes events. In DOLCE top level categories (Gangemi et al.[19]), Events are under the class of Perdurant/Occurrence. It seems to be controversial what the identity condition for events should be. Davidson [20] proposes a condition such that "events are identical if and only if they have exactly the same causes and effects". In any case it should be reasonable to assume that this class itself does not supply ICs but inherits them from the upper level classes.

*3 What we consider ICs for this class is as follows: Two instances of diseases are identical iff the two are experienced by the same host at the same time, are caused by the same agent (e.g. H5N1 virus for "H5N1 avian influenza") and have the same set of characteristic alterations/symptoms (e.g. inflammation of the lung for "pneumonia").

***Table 3:*** *Classification of concepts*

necessary IC: $E(x, t) \wedge \phi(x, t) \wedge E(x, t') \wedge \phi(y, t') \wedge x=y \rightarrow \Gamma(x, y, t, t')$
sufficient IC: $E(x, t) \wedge \phi(x, t) \wedge E(x, t') \wedge \phi(y, t') \wedge \Gamma(x, y, t, t') \rightarrow x=y$
(E : "actually exist at time t")

**Any property $\phi$ carries an IC (+I)** iff it is subsumed by a property supplying that IC.
**A property $\phi$ supplies an IC (+O)** iff i) it is rigid; ii) there is a necessary or sufficient IC for it; and iii) the same IC is not carried by all the properties subsuming $\phi$.

**<Dependency>** ([10], p.7)
**externally dependent property $\phi$ (+D):**
$\forall x \Box(\phi(x) \rightarrow \exists y\ \omega(y) \wedge \neg P(y, x) \wedge \neg C(y, x))$
(P: "is a part of")
(C: "is a constituent of")

Classification results are shown in Table 3. Most concepts such as ANATOMY, NON_HUMAN, and PERSON are classified as Type, whereas the concepts which were problematic in the first experiment were classified as Role: TRANSMISSION (Formal Role) and CASE (Material Role). According to the further classification of non-rigid concepts by Kaneiwa and Mizoguchi [18], these cases are classified as time-dependent concepts.

### 5.3 Modification of the schema

For some of the roles in Table 3, we modified their status in the annotation schema.

**CASE**

CASE and PERSON were problematic since we distinguished them according to the form of expression (unnamed/named), in addition to the case/non-case distinction. In order to cover the mentions which could not be annotated in the first experiment, we extended the scope of the PERSON class to include person instances in general, and eliminate the unnamed/named and case/non-case distinctions. We modified the annotation schema so that CASE is not the value of *cl* attribute, but is the *case* attribute which applies to the referred instance of PERSON. This attribute takes the value *true* when the mentioned instance is a confirmed case of disease,

*false* when the instance is not a case, and *putative* when the instance is a suspected case. Named case mentions and suspected case mentions are annotated as follows:

E9 Tests carried out in a UK laboratory confirmed that <NAME cl="PERSON" case="true">**M.A**</NAME>...

E10 <NAME cl="PERSON" case="putative">**a Taiwanese**</NAME> is suspected to have died of SARS

The meaning of *case* attribute-value pairs can be described in logical description and natural language as follows:

<...cl="PERSON" case="true">John</...>: **case(j)**
"It is true that the person **j** mentioned by "John" is an instance of the CASE class"

<...cl="PERSON" case="false">John</...>: **¬case(j)**
"It is false that the person **j** mentioned by "John" is an instance of the CASE class"

<...cl="PERSON" case="putative">John</..>:
**◇case(j)**
"It is possible that the person **j** mentioned by "John" is an instance of the CASE class"

As shown above, the values of the *case* attribute correspond to logical operators such as ¬ and ◇. The values of *case* attributes specify the modes of linkage between the referred concept and the CASE class. The formal basis we had in mind when formulating the *case* attribute are as follows: 1) every instance of a non-rigid class must be an instance of some rigid class, 2) the relations between a non-rigid class and its instance are often modified by modal/temporal operators. The first point drove us to create the case attribute which apply to instances of some rigid class, here, PERSON. The second point is the motivation for us to set values to include negative and modal operators. This schema can be extended if we allow a wider value range for the case attribute to include other modal/temporal operators, although currently we restrict the values to the three above.

It is worth noting that there is a trade-off between this revised schema and the former schema which is that we have increased the number of the markable entities, since we need to annotate unnamed, non-case mentions which are not directly related to the purpose of the system.

**TRANSMISSION**

We defined the *transmission* attribute which applies to mentions of ANATOMY, PRODUCT, PERSON and NON_HUMAN classes. As shown in the following examples, 'birds' are always related to NON_HUMAN, and take a '*true*' value only when they are mentioned as a source of infection. It can also take a '*putative*' value to cover mentions to possible sources of infection.

E11 Victims contract the virus from close contact with infected <NAME cl="NON_HUMAN transmission="true">**birds**</NAME>

**T_CHEMICAL /NT_CHEMICAL**

Concept classification revealed that T_CHEMICAL and NT_CHEMICAL have "the situation dependency obtained from extending types" discussed in [18] and have the same status as 'weapon' and 'table'. T_CHEMICAL includes chemicals mentioned as drugs in any context and those regarded as drugs in some context. Here we removed the two classes and made the parent node CHEMICAL as a class for annotation.

We then defined *therapeutic* attribute which applies to mentions of CHEMICAL and takes the value *true* when the entity is intended for therapeutic use and *false* otherwise.

As a result of the modifications above, our revised ontology is shown in Figure 2. We also added new classes CONDITION (status of patients: 'hospitalized' 'died 'in critical condition', etc) and OUTBREAK (collective disease incident: 'outbreak', 'pandemic', etc). Information about CONDITION is important for experts to know the rate of hospitalization and death and determine the alert level. Mentions of OUTBREAK include expressions which are specific to disease outbreak news, increasing the specificity of our detection system. We located PERSON and NON_HUMAN under metazoa, and added a *number* attribute (which takes *one* or *many* as its value) to be applied to PERSON instances.

With insights from the revised ontology we also changed the annotation method by dividing the process into two distinct stages as shown in Figure 3: 1) annotation of mentions to non-role (rigid) concepts and 2) annotation of role (non-rigid) concepts.

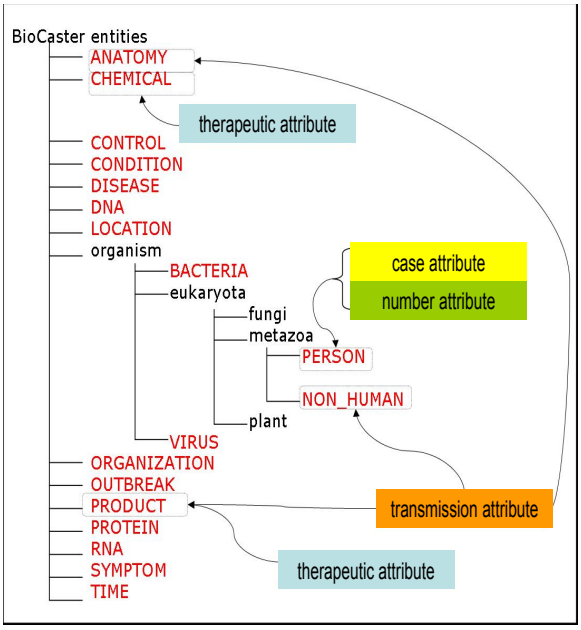


*Figure 2 Current ontology (simplified)*

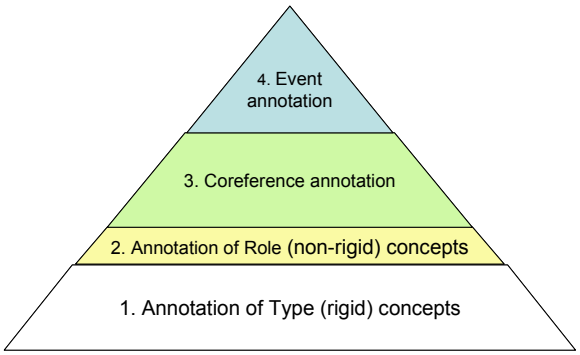


*Figure 3 Annotation schedule*

### 5.4 Results of annotation and NE recognizer training

We asked three PhD students to annotate a further 300 news articles. This time we used the revised annotation method 1 and 2 shown in Figure 3.

As a result of distinguishing between Role concepts (case, transmission, therapeutic) from others in the annotation schema, problem reports on these classes were reduced, and the annotation results were also improved. Contrary to our expectations, the complexity of the new annotation schema and the increased number of markable mentions seemed to have no negative influence on the annotator's speed.

The improvement can be seen empirically in the NER results. We re-annotated the corpus used in the first experiment using the revised annotation schema. This time the F-score for all classes rose to 79.96 (+3 compared to the previous result). Especially, significant increases of the F score were observed in the classes for PERSON (66.28; +11.33 compared to the previous result), case mentions among PERSON (65.63; +12.46), and NON_HUMAN (73.21; +5.21).

### 5.5 Remaining issues

Some of the problems reported in this second experiment were related to context dependency (anti-rigidity, situation dependency) discussed in Section 6.2.

The most difficult class seemed to be CONTROL (control measures to lower the risk of diseases). As shown in Table 3, we consider this class is also non-rigid, and it includes mentions which refer to subclasses of the CONTROL class regardless of situation ("quarantine" "vaccination"), and others which can be a control measure depending on the situation ("warning" "blockade"). This characteristic seems to cause the difficulty.

So far we have resolved the complexity of non-rigid concepts by defining attributes which apply to instances of rigid classes (e.g. the *case* attribute for the class PERSON). This strategy, however does not seem to be effective for CONTROL since it is not easy to identify a rigid superclass for CONTROL which can be realistically annotated in the text. For example, EVENT can be considered as a rigid class subsuming CONTROL, but currently it is not realistic to manually annotate every mention of an event. Currently we are seeking for a way to deal with this problem.

## 6. CONCLUSION

The study in this paper was motivated by our need for a high quality annotation schema to support detection of novel entities in the infectious disease outbreak domain. We discussed two experiments based on alternative approaches for constructing an ontology-based annotation schema. The amount of data in our study is relatively small but empirical results indicate support for our view that there is a positive effect in adopting well founded ontological principals over an ad-hoc task-based approach. Although this study is not a formal evaluation of ontologies, it is still an evaluation from the viewpoint of ontology application to the task of natural language annotation. The classification method of Guarino and Welty ([9], [10]) which was originally proposed to achieve consistency in the configurational structure of ontologies, was adapted and found to be useful for improving annotation performance.

An alternative possibility exists which we have not addressed in this paper which is to reformulate the tradition NER task to allow for overlapping (nested) and multi-class entities. This however introduces

significant additional complications in both the recognizer models and in the annotation schema so we have adopted a less radical formulation in this work.

As the next step in this study, we are now extending our simple taxonomy to a multi-lingual ontology; enriching the current taxonomic structure with domain-sensitive relations. The resulting ontology will be freely available for re-use. At the initial stage we are focusing on English, Japanese, Vietnamese, Thai, Chinese (standard) and Korean. We hope to add other Asia-Pacific languages in the future.

## Acknowledgements

We gratefully acknowledge partial funding support from the Japan Society for the Promotion of Science (grant no. 18049071). We also thank the anonymous reviewers for helpful comments.

## References

1. Ferguson NM, Cummings DA, Cauchemez S, Fraser C, Riley S, et al. Strategies for containing an emerging influenza pandemic in Southeast Asia. Nature 437: 209–214. 2005.
2. Grishman R, Huttunen S, and Yangarber R. Information extraction for enhanced access to disease outbreak reports. Journal of Biomedical Informatics, Vol. 35, No. 4, 236 - 246, 2002.
3. Public Health Agency of Canada. GPHIN system. http://www.phac-aspc.gc.ca/media/nr-rp/2004/2004_gphin-rmispbk_e.html
4. Aronson A.R. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proceedings of AMIA Symposium, 17–21, 2001.
5. Rindflesch T.C., Tanabe L, Weinstein J.N. and Hunter L. EDGAR: extraction of drugs, genes and relations from the biomedical literature. Proceedings of Pacific Symposium on Biocomputing 5:514-525, 2000.
6. Kim J.D., Ohta T, Tsuruoka Y, Tateishi Y, Collier N. Introduction to the Bio-entity Recognition Task of the JNLPBA workshop. Proceedings of the JNPBA, 70-76, 2004.
7. Yeh A, Morgan A, Colosimo M, Hirschman L. BioCreAtIvE task 1A: gene mention finding evaluation. BMC Bioinformatics 2005, 6(Suppl 1):S2.
8. Sowa J.F. Conceptual structures: Information processing in mind and machine. Addison-Wesley, New York; 1984.
9. Guarino N, Welty C. A formal ontology of properties. Dieng R, Corby O (eds.) Proceedings of EKAW-2000: The 12th International Conference on Knowledge Engineering and Knowledge Management, volume 1937: 97-112.
10. Guarino N, Welty C. Ontological analysis of taxonomic relations. Lander A, Storey V (eds.) Proceedings of ER-2000: The International Conference on Conceptual Modeling, vol. 1920, 210-224, Springer Verlag LNCS, Berlin, Germany.
11. Steimann F. On the representation of roles in object-oriented and conceptual modelling. Data and Knowledge Engineering35, 1: 83-106. 2000.
12. U.S. National Library of Medicine. Medical Subject Headings (MeSH), 2006.
13. Kim J.D., Ohta T, Tateishi Y, Tsujii J. GENIA corpus - a semantically annotated corpus for bio-textmining. Bioinformatics 19(suppl. 1), pp. i180-i182, Oxford University Press, 2003.
14. Hirschman L, Chinchor N. MUC-7 named entity task definition. Proceedings of the 7th Message Understanding Conference (MUC-7).
15. Hirschman L, Chinchor N, Grishman R, Sundheim B. Hub-4 Event Guidelines Version 2.6. http://www-nlpir.nist.gov/related_projects/muc/proceedings/hub4/guidelines.html
16. Vapnik, V. N. The Nature of Statistical Learning Theory, Springer-Verlag, New York, 1995.
17. Takeuchi, K and Collier, N. "Bio-medical entity extraction using support vector machines", in vol. 33, no.2, Artificial Intelligence in Medicine, Elsevier, pp. 125-137, 2005.
18. Kaneiwa K, Mizoguchi, R. An order-sorted quantified modal logic for meta-ontology. Proc. of the International Conference on Automated Reasoning with Analytic Tableaux and Related Methods (TABLEAUX 2005), Koblenz, Germany: 169-184, 2005.
19. Gangemi A, Guarino N, Masolo C, Oltramari A, Schneider L. Sweetening ontologies with DOLCE. Benjamins et al. (eds.), Proceedings of the 13th European Conference on Knowledge Engineering and Knowledge Management (EKAW2002), 166-181, Sigüenza, Spain, 2002.
20. Davidson D. The Individuation of events. Rescher N (ed) Essays in Honor of Carl G. Hempel: 216-234, 1969, D. Reidel.