Applied Ontology 4 (2009) 5–20
DOI 10.3233/AO-2009-0062
IOS Press

The development of a schema for semantic annotation: Gain brought by a formal ontological method

Ai Kawazoe [a,*], Lihua Jin [a], Mika Shigematsu [b], Daisuke Bekki [c], Roberto Barrero [d], Kiyosu Taniguchi [b] and Nigel Collier [a]

[a] *National Institute of Informatics, Tokyo, Japan*
[b] *National Institute of Infectious Diseases, Tokyo, Japan*
[c] *Ochanomizu University, Tokyo, Japan*
[d] *Centre for Comparative Genomics, Murdoch University, Perth, WA, Australia*

Abstract. In this paper, we will report annotation experiments which show the advantage of applying a formal ontological methodology for constructing a schema for semantic annotation to mark up terms in the public health domain. We demonstrate that (1) a traditional task-oriented approach with a simple schema can cause several critical problems, and (2) the performance of annotators and the quality of annotated corpus is improved by applying formal ontological methodology in analyzing 'markable' categories of concepts and restructuring the schema. These results show that disciplined methods are useful for controlling the development of even quite modest semantic structures like annotation schema for entity recognition. We also report philosophical/logical considerations and decisions we made when we adopted the formal approach.

Keywords: Semantic annotation, ontology, named entity recognition, meta-property, text mining

1. Introduction

Semantic annotation of named entities (NE) is a basic technology for Information Extraction (IE), which aims to extract structured information from unstructured natural language texts. Much previous work in IE has relied on high quality human annotated training data for constructing and evaluating named entity recognizers (NERs) (e.g., Kim et al., 2003; Franzén et al., 2002; Tanabe et al., 2005). We can reasonably assume that any inconsistency introduced by ontological inconsistencies into the annotation schema (a specification of an annotation task, which includes definitions of 'markable' categories, tags and attributes used for annotation) should be harmful for annotation performance, by both humans and machines. However, it is not easy to provide human annotators with a well-defined, comprehensive schema for semantic annotation. The difficulty in defining categories for annotation ('markable' or 'target' categories) had been recognized already when the target was only the seven 'standard' categories of person, location, organization, and several numeric expressions in the 1990's Message Understanding Conference (Grishman & Sundheim, 1996). With the increasing use of NE in domains such as biology

*Corresponding author: Ai Kawazoe, National Institute of Informatics, Hitotsubashi 2-1-2, Chiyoda-ku Tokyo 101-8430, Japan. Tel.: +81 3 4212 2536; Fax: +81 3 4212 2536; E-mail: zoeai@nii.ac.jp.

and medicine, recent extensions of markable categories such as organisms and biological substances make the problem more serious.

In this paper, we will report annotation experiments which show the advantage of applying a formal ontological methodology for constructing an annotation schema to mark up terms in the public health domain. We are developing BioCaster, a text mining-based system for infectious disease detection and tracking, whose key feature is the use of automated learning methods to identify novel entities and events using features derived from examples by human annotators. In our early development of BioCaster, it became clear that we needed a rigorous schema for markable entities. Surprisingly, while there have been several studies on the mapping problem between terms and coding systems such as the UMLS Metathesaurus (Aronson, 2001) as well as biomedical annotation experiments (e.g., Rindflesch et al., 2000; Kim et al., 2004; Yeh et al., 2005), there have been to the best of our knowledge few studies (e.g., Bouaud et al., 1998) 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 poorly-considered markable categories can indeed be harmful to accuracy, and that re-organizing the schema by analyzing markable categories with formal meta-properties (Guarino & Welty, 2000a, 2000b, and related studies) produced better results, despite the added complexity. The results will show the benefit of having formal tools to construct a simple schema for NE semantic annotation. We also describe the philosophical/logical considerations that we made when applying meta-properties to our categories of interest.

The rest of this paper is organized as follows. In Section 2, we introduce the purpose of BioCaster and describe the needs of epidemiologists as primary users of the system. We present an overview of the task-oriented, simple annotation schema we originally adopted (Section 3), and report several problems that occurred in the annotation experiment (Section 4). We describe the analysis of our markable categories and the reorganization of the annotation schema in Section 5, and report the result of the second annotation experiment in Section 6.

2. The Biocaster system

2.1. Motivation

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 early has clear implications for morbidity and mortality, as well as the feasibility of containment (Ferguson et al., 2005). However, the lack of a 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 relying on 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 (Grishman et al., 2002) but only a few systems are currently actively deployed, the most prominent of which being the Global Public Health Intelligence Network (GPHIN, Public Health Agency of Canada), a successful but semi-closed system used by the WHO. BioCaster is a text mining system based on an openly available multilingual ontology for proactive notification of priority disease outbreaks. We are developing the core text mining module by using machine learning from multilingual collection of news articles annotated by humans. The initial target languages are English, Japanese, Vietnamese and Thai.

2.2. *Target categories for information extraction*

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 in Japan 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 integrate an additional level of detail about the host which includes genes and their products, identified

Table 1
Categorization of concepts

Focus	Description	Example properties	Concept types
Agent	Pathogen	Infectivity, pathogenicity, virulence, incubation period, communicability	VIRUS, BACTERIA, PARASITE*, FUNGI*
Transmission	The delivery or dispersal method	Dermal, oral, respiratory	TRANSMISSION
Host	Person 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.

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

Table 2
List of important concepts

Classes	Examples	Description
ANATOMY	*Liver, pancreas, nervous system, HeLa cell*	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[a]
T_CHEMICAL	*Relenza, immunosuppressive drug, oseltamivir*	Chemicals intended for the treatment of diseases[a]
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 DNA groups, families, molecules, domains and regions[b]
LOCATION	*Viet Nam, Jakarta, Sumatra Island, Asia*	A politically or geographically defined location[c]
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[c]
PERSON	*Jean Chretien, Murray McQuigge*	A named person or family
PRODUCT	*Botulism antitoxin, Influenza vaccine*	Biological product (e.g. vaccines, immune serums)
PROTEIN	*STAT, RNA polymerase II alpha subunit*	Includes the names of proteins, groups, families, molecules, complexes and substructures[b]
RNA	*IL-2R alpha transcripts, TNF mRNA*	Includes the names of RNA groups, families, molecules, domains and regions[b]
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[d]
TRANSMISSION	*HIV-tainted blood products, BSE-infected cows*	Source of infection
VIRUS	*Ebola virus, HIV*	Viruses such as HIV, HTLV, EBV[b]

Note: Descriptions marked with [a], [b], [c], [d] are based on those in MeSH (U.S. National Library of Medicine, 2006), GENIA ontology (Kim et al., 2003), MUC-7 (Hirschman & Chinchor, 1997), and HUB-4 (Hirschman et al., 2005), respectively.

with a [§] in Table 1. Finally we had 19 categories of concepts which we want to identify in news texts (Table 2).

3. The original annotation schema

At this stage we were aware that some of the important concepts in Table 2 are intrinsically different from other concepts. For example, CASE and TRANSMISSION represent roles (discussed in Sowa, 1984; Guarino & Welty, 2000a, 2000b; Steimann, 2000 among others) and are dependent on the existence of events in which they participate, while most others, such as PERSON, BACTERIA, and NON_HUMAN are not.

However, the first approach we adopted for constructing an annotation schema was rather task-oriented, as is the case in many traditional IE studies. We did not make any distinctions between role and

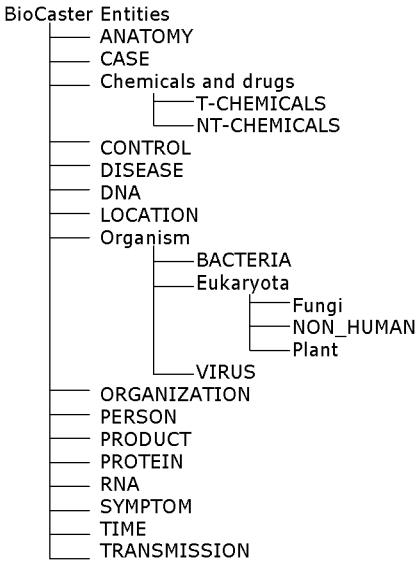

Fig. 1. Initial markable classes.

non-role concepts. We constructed a hierarchical tree of markable categories shown in Fig. 1 (markable categories are in capital letters). Note that role concepts such as CASE and TRANSMISSION have the same status as others. All categories are represented as classes (unary universals). We considered that all classes are disjoint and no entity belongs to more than one class, following many of the previous NER works.

By adopting the task-based approach, we expected that the corresponding annotation schema would be uncomplicated, since instances of role classes such as CASE are annotated in the same way as those of other classes, e.g. PERSON:

<NAME cl="PERSON">Kofi Annan</NAME>

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

One advantage of this task-oriented approach is that we can annotate exactly what must be contained in the event frame (a frame to represent extracted information of an event and its arguments in a structured way). For example, we can exclude from annotation non-named, non-case mentions, which do not play a role in the event frame. 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. 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 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 "*related*" is to cover two types of relations between mentioned entities and the class hierarchy we want to describe. The first one is "is an instance of" and the other "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 usually refer to particulars, whereas names of chemical substance, viruses and proteins often refer to universals. This is one of the factors which make ontology-based annotation a complicated process. It should be noted though that we intend to work towards a clear distinction between the two types of relations in future work.

4. Annotation experiment 1

We developed annotation guidelines to mark up non-overlapping mentions of markable classes, and hired two graduate students (MSc in Informatics) as annotators. In the guidelines, we instructed annotators to markup only the single most appropriate class for a markable text, and prohibited multiple classes. After one 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.

4.1. Annotation results and problems

During the first annotation experiment, we had many problem reports from 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 annotation schema. Follow-up analysis on the corpus yielded the following results, indicative of errors in the schema:

- Gaps in the annotation schema shown by the existence of mentions of entities which we wanted to annotate, but were not covered by the annotation schema.
- Ambiguity between role concepts and other concepts.
- Idiosyncratic annotations which are forced on annotators due to the disjointness between classes in the schema.

4.1.1. Gaps in the annotation schema

At the initial stage of our analysis, we considered that distinction between CASE (confirmed cases of a disease which are unnamed humans) and 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 number of 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]

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

In addition, we found that there were more frequent mentions of putative cases than we had expected. These mentions ("epistemology-loaded terms" discussed in Bodenreider et al., 2004) were often annotated as CASE by annotators although we restricted the scope of this class only to confirmed cases:

E2 **A suspected case of SARS** is being investigated.

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.

4.1.2. Ambiguity caused by role 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 '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.

4.1.3. Idiosyncratic annotations due to the disjointness between classes in the schema

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. However, because of the principle of disjoint classes, the annotator was forced to select only one class.

4.2. Empirical results from training an NE recognizer

We trained a support vector machine (Vapnik, 1995; for details, see Takeuchi & Collier, 2005) for NE recognition based on an 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. The problematic classes included PERSON, CASE and NON_HUMAN (many instances of which had ambiguity with TRANSMISSION). These classes had F-scores below average: PERSON (54.95), CASE (53.17), NON_HUMAN (68.0).

[3] Available from: http://cl.aist-nara.ac.jp/∼taku-ku/software/TinySVM.

5. Application of a formal methodology

5.1. Formal analysis of markable categories

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 role and non-role classes. We decided to take an alternative, formal approach, and distinguish role concepts from other concepts in the annotation schema.

The first step was to analyze our markable categories by using three meta-properties (rigidity, identity, dependency) proposed by Guarino & Welty (2000a, 2000b), which are the basis of the ONTOCLEAN methodology (Guarino & Welty, 2002) developed for 'cleaning up' the hierarchical structure of ontologies. Definitions of the meta-properties we used are as follows:

<Rigidity> (Guarino & Welty, 2000a, p. 4)

rigid property ϕ(+R): $\forall x \phi(x) \rightarrow \Box \phi(x)$

anti-rigid property ϕ($\sim$R): $\forall x \phi(x) \rightarrow \neg \Box \phi(x)$

($\Box$: Necessity Operator)

<Identity> (Guarino & Welty, 2000a, p. 5)

Identity Condition (IC): An identity condition is a formula Γ that satisfies either of the following:[4]

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 exists at time t")

Any property ϕ carries an IC (+I) iff it is subsumed by a property supplying that IC.

A property ϕ 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 ϕ.

<Dependency> (Guarino & Welty, 2000a, p. 7)

externally dependent property ϕ(+D):

$\forall x \Box(\phi(x) \rightarrow \exists y \psi(y) \wedge \neg P(y,x) \wedge \neg C(y,x))$

(P: "is a part of", C: "is a constituent of", ψ is another property).

In order to interpret these definitions in modal logic properly, we referred to one of the related works, *WonderWeb Deliverable D17* (Masolo et al., 2002), for the underlying logical/philosophical assumptions. According to this work, they assume:

- Modal logic system S5, with Barcan Formula (BF).
- Possibilist view of Lewis (1983), a kind of modal realism: non-actual, but possible entities are included in the domain.
- Eternalist view of time: all past, present and future entities and intervals are included in the domain.

[4] In Guarino & Welty (2000b), 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.

However, in the process of analyzing our own markable categories, we found that some additional philosophical/logical considerations and practical settings are required in order to make an epistemologically-plausible basis for the analysis. We will describe these below.

5.1.1. On possible worlds

As shown in its definition above, rigidity seems to be a convenient meta-property to distinguish between role concepts (e.g., CASE) and non-role concepts (e.g., PERSON) among our markable categories. However, when we judge the rigidity of a particular category, we need to decide the scope of 'accessible' possible worlds, which are quantified over by the necessity operator. It is obvious that for our purpose it is not desirable to include every world we can think of or talk about. If we count fictional worlds as accessible from our world, rigid properties may become nonexistent.

The strategy we took here is to restrict ourselves to only temporal/situational interpretation of possible worlds (i.e., we focus on W_{Tim} and W_{Sit} in Kaneiwa & Mizoguchi, 2005). This makes 'transworld identity' a trivial, intuitive notion, and has advantage over the alternatives which involve counterfactuals and fictions, from a practical point of view.

5.1.2. On events in modal predicate logic

Some of our markable categories include event categories. The characterization of events as ontological individuals dates back to Davidson (1967), and recently a neo-Davidsonian theory of events has been introduced (Parsons, 1990). Such characterization enables us to apply meta-properties to properties of events. However, we cannot simply incorporate the (neo-)Davidsonian events into the modal logic system with BF which we assume following Masolo et al. (2002), because of the problem stated below:

(BF) $\quad \forall x \Box \phi(x) \rightarrow \Box \forall x \phi(x).$

As discussed in Hughes and Cresswell (1996, Chapter 16), from a philosophical point of view, the plausibility of adding BF has been controversial, since a naive interpretation of BF reflects a view that every possible world shares the same domain of entities. When applied to Davidsonian event semantics, this interpretation of BF causes a problem, since it will allow an event that exists at some time point to exist in all the time points. For example, if there is an event of John's marriage, the neo-Davidsonian style formula '$\exists e$ (marry(e) $\wedge$ agent (e, John))' will always be true. This means that the modal logical system cannot function as temporal logic.

In order to avoid the philosophical problem in BF, Hughes & Cresswell (1996) introduced the definition of a universal quantifier ('actualist' quantifier Π) with the existence predicate (E). We assume that Guarino & Welty (2000a, 2000b) and related works also adopt this strategy. The validity condition for E is stated as follows (Hughes & Cresswell, 1996, p. 292):

(VE) $\quad \langle u, w \rangle \in V(E)$ iff $u \in D_w \quad (D_w$ is the domain of quantification of $w)$.

By the existence predicate E, we can describe which objects are actual in a particular world, and thus we can restrict the domain of universal quantifier to only actual objects, excluding non-actual, possible objects from the domain of quantification. The 'actualist' universal quantifier Π is defined as follows:

$$\Pi x \phi(x) \equiv_{\text{def}} \forall x (E(x) \rightarrow \phi(x)).$$

This enables us to maintain BF as it is, accommodating our intuition about 'universal quantification' to the quantification over actual objects.

In order to avoid the problem with (neo-)Davidsonian events stated above, we can define 'actualist' version of existential quantifier (Σ) as follows:

$$\Sigma x\phi(x) \equiv_{\text{def}} \exists x(E(x) \wedge \phi(x)).$$

With this existential quantifier, we can rewrite the neo-Davidsonian formula which expresses the occurrence of John's marriage as follows:

$$\Sigma e(\text{marry}(e) \wedge \text{agent}(e, \text{John})) \equiv \exists e(E(e) \wedge \text{marry}(e) \wedge \text{agent}(e, \text{John})).$$

This formula becomes false in the worlds (including time points) where John's marriage does not occur. With this settings we can combine (neo-)Davidsonian event semantics with BF, and make a basis for analyzing properties of events.

Based on the considerations and settings above, we analyzed our markable categories, and the 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).

Table 3
Classification of concepts

	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	−O[a]	+I	+D	Material role
DISEASE	+R	+O[b]	+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

[a]This class includes events. In DOLCE top level categories (Gangemi et al., 2002), events are under the class of perdurant/occurrence. It seems to be controversial what the identity condition for events should be. Davidson (1969) 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.

[b]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").

5.2. Modification of the schema

We modified the status of some role concepts in Table 3 and reconstructed the annotation schema as discussed below.

5.2.1. 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 suspected case of SARS**</NAME> is being investigated.

The meaning of *case* attribute-value pairs can be expressed in formal notation and natural language as follows:

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

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

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

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 role. The formal bases we had in mind when formulating the *case* attribute includes the following elements: (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 a rigid class, here, PERSON, to indicate whether they are also instances of a non-rigid subclass, i.e., a class of persons which has the role CASE, or not. 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 (true, false and putative).

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.

5.2.2. 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>

5.2.3. T_CHEMICAL /NT_CHEMICAL

Concept classification revealed that T_CHEMICAL and NT_CHEMICAL have "the situation dependency obtained from extending types" discussed in Kaneiwa & Mizoguchi (2005) 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 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.

The resulting schema was more complex than the task-based schema due to the fact that role concepts have a different status compared to other concepts, i.e., they are annotated in different ways. In order to achieve ontological consistency we also need to annotate more mentions than in the former approach, including those that will not instantiate event frames.

As a result of the modifications described above, our revised markable class hierarchy is shown in Fig. 2. We also added the 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 un-

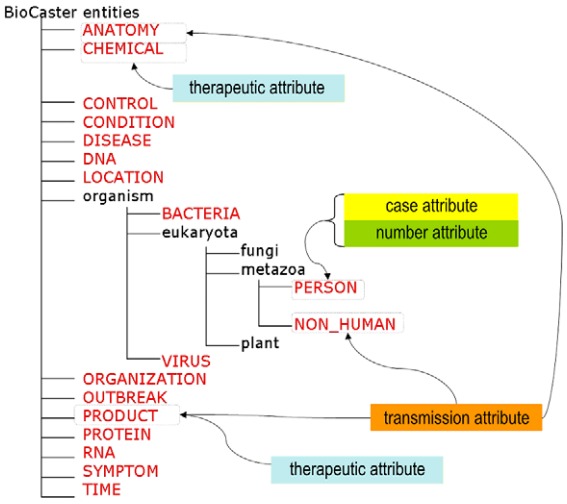

Fig. 2. Current markable classes and attributes.

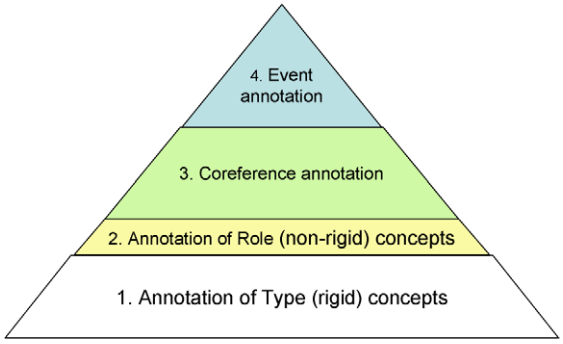

Fig. 3. Annotation schedule.

der their common parent concept metazoa, and added a *number* attribute (which takes *one* or *many* as its value) to be applied to PERSON instances.

In order to reflect changes in the revised ontology, we also changed the annotation method by dividing the process into two distinct stages as shown in Fig. 3: (1) annotation of mentions to non-role (rigid) concepts and (2) annotation of role (non-rigid) concepts.

6. Annotation experiment 2

6.1. Results of annotation and NE recognizer training

We asked three PhD students to annotate a new set of 300 news articles. This time we used the revised annotation methods 1 and 2 shown in Fig. 3.

As a result of distinguishing Role concepts (case, transmission, therapeutic) from other concepts in the annotation schema, problem reports were less frequent, and the annotation results 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). Analysis revealed that the reduction in errors seemed to come mainly from differentiating between different concept types such as rigid and non-rigid. In retrospect this became clearer as the features needed to automatically label non-rigid concepts should be more complex than those for rigid concepts. For example, the decision about whether or not to annotate 'the 34-year old man' as CASE depends on a cue word like 'patient' in the next sentence or even the next paragraph.

6.2. Remaining issues

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

The most difficult class to annotate to seemed to be CONTROL (control measures to lower the risk of diseases). As shown in Table 3, we consider this class to be 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 super class 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. Further research is needed to address this problem.

7. Conclusion

The study in this paper was motivated by our need for a high-quality annotation schema to support the detection of novel entities in the infectious disease outbreak domain. We discussed two experiments based on alternative approaches for constructing an 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 methodologies such as Guarino & Welty (2000a, 2000b) over an *ad hoc* task-based approach. Through the discussion of the two experiments, we have shown that formal tools are useful to guide the development of even quite modest (in terms of size) structures such as schema for entity annotation.

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. It should be emphasized that the positive effect of the formal methodology was empirically assessed not only by the performance of human annotators, but also by the results of NER training. This indicates the potential for evaluating different models of knowledge, including annotation schemas and domain ontologies, using NLP applications.

An alternative, not addressed in this paper, is to reformulate the traditional 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 schemata, so we have adopted a less radical formulation in this work.

One of the issues not yet addressed in our annotation work is systematic polysemy. We observed confusion in both annotator reports and annotation results between annotations to LOCATION and ORGANIZATION, for some terms including country names such as 'Japan'. As suggested by e.g., Gangemi et al. (2000), polysemy is another issue where formal methodologies for conceptual analysis offer a potential solution. In future work, we will attempt to apply formal methodologies to improve the annotation results for polysemous expressions.

As the next step in this study, we are now extending our simple taxonomy to a multi-lingual ontology; and we are enriching the current taxonomic structure with domain-specific relations such as causation relations between pathogens and diseases (Collier et al., 2007). The latest version of BioCaster ontology is available online at http://biocaster.nii.ac.jp/. 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 two reviewers of Applied Ontology for their valuable comments.

References

Aronson, A.R. (2001). Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In *Proceedings of AMIA Symposium 2001*, Washington, DC (pp. 17–21).

Bodenreider, O., Smith, B. & Burgun, A. (2004). The ontology–epistemology divide: A case study in medical terminology. In *Proceedings of the Third International Conference on Formal Ontology in Information Systems (FOIS 2004)*, Torino, Italy (pp. 185–195).

Bouaud, J., Bachimont, B., Charlet, J. & Zweigenbaum, P. (1995). Methodological Principles for Structuring an "Ontology". In *Proceedings of the IJCAI'95 Workshop on Basic Ontological Issues in Knowledge Sharing*, Montreal, Canada (pp. 95–148).

Collier, N., Kawazoe, A., Jin, L., Shigematsu, M., Dien, D., Barrero, R. et al. (2007). A multilingual ontology for infectious disease surveillance: rationale, design and challenges. *Language Resource and Evaluation*, *40*(3/4), 405–413.

Davidson, D. (1967). The logical form of action sentences. In N. Rescher (ed.), *The Logic of Decision and Action*. Pittsburgh, PA: University of Pittsburgh Press.

Davidson, D. (1969). The individuation of events. In N. Rescher (ed.), *Essays in Honor of Carl G. Hempel* (pp. 216–234). Dordrecht: D. Reidel.

Ferguson, N.M., Cummings, D.A., Cauchemez, S., Fraser, C., Riley, S., Meeyai, A. et al. (2005). Strategies for containing an emerging influenza pandemic in Southeast Asia. *Nature*, *437*, 209–214.

Franzén, K., Eriksson, G., Olsson, F., Asker, L., Lidén, P. & Cöster, J. (2002). Protein names and how to find them. *International Journal of Medical Informatics*, *67*(1–3), 49–61.

Gangemi, A., Pisanelli, D.M. & Steve, G. (2000). Understanding systematic conceptual structures in polysemous medical terms. In *Proceedings of the 2000 AMIA Fall Symposium*, Los Angeles, CA (pp. 285–289).

Gangemi, A., Guarino, N., Masolo, C., Oltramari, A. & Schneider, L. (2002). Sweetening ontologies with DOLCE. In *Proceedings of the 13th European Conference on Knowledge Engineering and Knowledge Management (EKAW2002)*, Siguenza, Spain (pp. 166–181).

Grishman, R. & Sundheim, B. (1996). Message Understanding Conference-6: A brief history. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING)*, Copenhagen, Denmark (pp. 466–471).

Grishman, R., Huttunen, S. & Yangarber, R. (2002). Information extraction for enhanced access to disease outbreak reports. *Journal of Biomedical Informatics*, *35*(4), 236–246.

Guarino, N. & Welty, C. (2000a). A formal ontology of properties. In *Proceedings of EKAW-2000: The 12th International Conference on Knowledge Engineering and Knowledge Management*, Juan-les-Pins, France (Vol. 1937, pp. 97–112).

Guarino, N. & Welty, C. (2000b). Ontological analysis of taxonomic relations. In *Proceedings of ER-2000: The International Conference on Conceptual Modeling*, Salt Lake City, UT (Vol. 1920, pp. 210–224).

Guarino, N. & Welty, C. (2002). Evaluating ontological decisions with OntoClean. *Communications of the ACM*, *45*(2), 61–65.

Hirschman, L. & Chinchor, N. (1997). MUC-7 named entity task definition. In *Proceedings of the 7th Message Understanding Conference (MUC-7)*, Fairfax, VA.

Hirschman, L., Chinchor, N., Grishman, R. & Sundheim, B. (2005). Hub-4 Event Guidelines Version 2.6. Available at: http://www-nlpir.nist.gov/related_projects/muc/proceedings/hub4/guidelines.html.

Hughes, G.E. & Cresswell, M.J. (1996). *A New Introduction to Modal Logic*. New York, NY: Routledge.

Kaneiwa, K. & Mizoguchi, R. (2005). An order-sorted quantified modal logic for meta-ontology. In *Proceedings of the International Conference on Automated Reasoning with Analytic Tableaux and Related Methods (TABLEAUX 2005)*, Koblenz, Germany (pp. 169–184).

Kim, J.D., Ohta, T., Tateishi, Y. & Tsujii, J. (2003). GENIA corpus – a semantically annotated corpus for bio-textmining. *Bioinformatics*, *19*(Suppl. 1), i180–i182.

Kim, J.D., Ohta, T., Tsuruoka, Y., Tateishi, Y. & Collier, N. (2004). Introduction to the Bio-entity recognition task of the JNLPBA workshop. In *Proceedings of the JNPBA*, Zeneva, Switzerland (pp. 70–76).

Lewis, D. (1983). New work for a theory of universals. *Australasian Journal of Philosophy*, *61*(4), 343–377. (Reprinted in Mellor, D.H. and Oliver, A. (eds), *Properties*, Oxford University Press, 1997.)

Masolo, C., Borgo, S., Gangemi, A., Guarino, N., Oltramari, A. & Schneider, L. (2002). *WonderWeb Deliverable D17*. The WonderWeb Library of Foundational Ontologies and the DOLCE ontology. Preliminary Report (ver. 2.0, 15-08-2002).

Parsons, T. (1990). *Events in the Semantics of English: A Study in Subatomic Semantics*. Cambridge, MA: MIT Press.

Public Health Agency of Canada. GPHIN system. Available at: http://www.phac-aspc.gc.ca/media/nr-rp/2004/2004_gphin-rmispbk_e.html.

Rindflesch, T.C., Tanabe, L., Weinstein, J.N. & Hunter, L. (2000). EDGAR: extraction of drugs, genes and relations from the biomedical literature. In *Proceedings of Pacific Symposium on Biocomputing*, Hawaii, HI (Vol. 5, pp. 514–525).

Sowa, J.F. (1984). *Conceptual structures: Information Processing in Mind and Machine*. New York: Addison-Wesley.

Steimann, F. (2000). On the representation of roles in object-oriented and conceptual modelling. *Data and Knowledge Engineering*, *35*(1), 83–106.

Takeuchi, K. & Collier, N. (2005). Bio-medical entity extraction using support vector machines. *Artificial Intelligence in Medicine*, *33*(2), 125–137.

Tanabe, L., Xie, N., Thom, L.H., Matten, W. & Wilbur, W.J. (2005). GENETAG: a tagged corpus for gene/protein named entity recognition. *BMC Bioinformatics*, *6*(Suppl. 1), S3.

U.S. National Library of Medicine (2006). Medical Subject Headings (MeSH).

Vapnik, V.N. (1995). *The Nature of Statistical Learning Theory*. New York: Springer-Verlag.

Yeh, A., Morgan, A., Colosimo, M. & Hirschman, L. (2005). BioCreAtIvE task 1A: gene mention finding evaluation. *BMC Bioinformatics*, *6*(Suppl. 1), S2.