Chapter 5
KYOTO: A Knowledge-Rich Approach to the Interoperable Mining of Events from Text

Piek Vossen, Eneko Agirre, German Rigau, and Aitor Soroa

Abstract To automatically understand text, a crucial step is to extract events and their participants. The same event can be *packaged* in many different ways in a language. Capturing all these ways with sufficient precision is a major challenge. This becomes even more complex, when we consider texts in different languages on the same topic. We describe a knowledge-rich event-mining system developed for the Asian-European project KYOTO that can extract events in a uniform and interoperable way, regardless of the way they are expressed and in which language. To achieve this, we developed an open text representation format, semantic processing modules and a central ontology that is shared across seven languages. We implemented a semantic tagging approach that performs off-line reasoning and a module for detecting semantic and linguistic patterns in the tagged data to extract events from a large variety of expressions. The system can efficiently handle large volumes of documents and is not restricted to a specific domain. We applied the system to an English text on estuaries.

5.1 Introduction

Information Extraction (IE) can be described as the task of filling template information from previously unseen text which belongs to a predefined domain [18]. Standard IE systems are based on language-specific pattern matching [13], where

P. Vossen (✉)
VU University Amsterdam, De Boelelaan 1105, 1081HV, Amsterdam, The Netherlands
e-mail: piek.vossen@vu.nl

E. Agirre · G. Rigau · A. Soroa
University of the Basque Country, M. de Lardizabal Pasealekua 1, 20018, Donostia, Spain
e-mail: e.agirre@ehu.es; german.rigau@ehu.es; a.soroa@ehu.es

A. Oltramari et al. (eds.), *New Trends of Research in Ontologies and Lexical Resources*, Theory and Applications of Natural Language Processing,
DOI 10.1007/978-3-642-31782-8_5, © Springer-Verlag Berlin Heidelberg 2013

each pattern consists of a regular expression and an associated mapping from syntactic to logical form. The use of ontologies in IE is an emerging field [3]: linking text instances with elements belonging to the ontology, instead of consulting flat gazetteers. IE can be considered as a knowledge-rich approach to filter information from text, mostly using very specific background models. They focus on satisfying precise, narrow, pre-specified requests (e.g. to extract *all directors of movies*) and are able to only detect precise matches (e.g. from web documents) while they do not need to understand the remainder of the text.

This approach does not extend well for event mining, since this latter problem demands complex analysis of different semantic components: the events, their participants and their semantic roles, that can be expressed in many different ways or left implicit. Furthermore, existing semantic paradigms for modeling events such as FrameNet [2] and TimeML [19] are built upon specifications of events that often contradict each other, and no unitary framework for the analysis of events, relations and event participants over time has been applied to document processing so far.

We present a knowledge-rich approach to mining events from text that can handle a large amount of expressions of event information and can be applied to many different languages. It uses an open text representation system and a central ontology that is shared across languages. Ontological implications are inserted in the text through off-line reasoning and ontological tagging. We built a flexible pattern-matching module that searches for ontological and shallow linguistic event structures defined through simple XML profiles. We show that a rich ontology linked to large vocabularies can be used to extract event data from a wide variety of expressions from different languages in an interoperable way. It represents a first step towards the semantic modeling of events in text on a large scale and involving a wide variety of deeper ontological knowledge. The system is developed in the Asian-European project KYOTO[1] and tested for the environment domain.

In the next section, we first explain in more detail the large variety of ways in which event-data can be packaged in languages. In Sect. 5.3, we describe the general architecture of the KYOTO system and in Sect. 5.4 the knowledge structure used. Sect. 5.5 explains the off-line reasoning and ontological tagging process. In Sect. 5.6, we describe the module for mining knowledge from the text that is enriched with ontological statements. Finally in Sect. 5.7, we describe the results of applying the system to text on environmental issues for large estuaries.

5.2 Packaging of Events

People use a large variety of ways to refer to events in language. Whereas *things* such as *fish* can only be referred to by nouns and names in most languages, words in any part of speech can refer to events, e.g. *migration* (noun), *migrate* (verb),

[1] www.kyoto-project.eu

migratory (adjective) or *The Migration Period* (named event). Consequently, event mentions in text exhibit a large variety of syntactic structures as illustrated by the following examples taken from the Internet (italics and bold face added):

- *Adjectival reference:*

  1. In Europe, most **migratory fish species** *completing their cycle between the sea and the river* are currently in danger.
  2. Dams, culverts and other barriers currently block *the movement* of **migratory fish** *to spawning grounds*.

- *Nominal reference:*

  3. Downstream **migration of juvenile fishes** is an adaptation aimed at finding *habitat* and new areas *for feeding*, thereby expanding the feeding areas of the species.
  4. Historically, local economies flourished from the *annual shad run in the spring*, when the **fishes' upriver migration** begins.
  5. Species such as salmon, sturgeon, lampreys and various Cyprinids all *have* **anadromous migration** *patterns*, while Eels *have* **catadromous migration** *patterns*.

- *Verbal reference:*

  6. **Eel migrate** in the opposite sense they spend the longest time of their life in the *river* and *spawn in the sea*.
  7. **Menhaden migrate** *into Chesapeake Bay*

A number of issues are illustrated by these examples. First of all, the syntactic structures vary widely and cannot easily be covered through patterns. In the case of adjectival usage the noun that it modifies (*fish*) is the participant doing the migration. In the case of nominal usage, it can be the following *of*-phrase that holds the participant but also the possessive construction (*fishes'*) that proceeds it. More extreme is the sentential construction in which participants *have* patterns of *migration*, from which the reader needs to infer that the *fish* actually participate in the event. In the case of the verbal expression of *migrate*, it is the subject noun that refers to the participant.

In addition to references to events and the participants, we also find references to other events that are somehow semantically related to *migration*. *Cycles between the sea and the river* (example 1) actually co-refer with the migration process, where species travel from *sea* to *river* and back, and fill in details. In other cases, reference is made to events that have an effect on *migration*, e.g. *barriers block* (example 2), or represent the reasons, e.g. *finding habitat and new areas for feeding* (example 3). We see here that the event *migration* is packaged in many different ways and that the sentence includes aspects of the events (italics phrases) that are either directly related to it (repeating the event and filling in other elements) or that have some causal relation to it.

In some cases the same event is referred to without using the word *migrate* or any of its derived forms. These are called conceptual references, as opposed to the previous lexical references:

- *Conceptual reference:*

    1. Some measures were taken in the late 1880s to provide access for anadromous fishes around dams by construction of rudimentary fishways, or by stocking fish into habitats that historically supported large runs.
    2. The allis shad used to be found in the large rivers but is now extinct in the Netherlands.

Only through our knowledge that *anadromous fish* is a type of *migratory fish* and *allis shad* is a type of *anadromous fish*, we can interpret the rest of these sentences in relation to the *fish migration* process.

Packaging of events is a well-known phenomenon in cognitive science and cognitive linguistics literature. For example, Majid et al. [14] argue that events in language are always packaged through the choice of semantic roles. Within computational approaches this is less commonly accepted as a starting point. A computer program that tries to reconstruct the *migration* event from any of these texts faces a major challenge. It not only needs to deal with the different syntactic structures but also needs to have access to knowledge about migration and decide on the interpretation of the different phrases in relation to the event. The above examples are all in English, but events could be extracted from text in different languages, requiring the following capabilities:

1. Handle a large variety of syntactic structures to express events and (causal) relations between events.
2. Have a semantic typing of the words in the text: what words refer to events and what words can refer to the participants.
3. Know what participants an event takes and what their roles are.
4. Have rich knowledge about the type of event or process to understand causal relations with other events and conditions.
5. Have a large and rich database of semantic relations to inherit properties to more specific words and concepts.
6. Use a uniform and interoperable approach across different languages.

To solve this problem completely, large amounts of deep background knowledge need to be paired with knowledge about the way reference can be made to events and participants in and across languages. In this article, we describe a first step in tackling these problems using a knowledge-rich approach that is interoperable across different languages. Our solution includes the following elements:

- The structure of text is represented in a uniform way across different languages.
- All textual elements are converted into ontological elements in the same way across these languages.
- We use an ontological model that is designed to model events and relations between events.

- The vocabularies of the different languages are mapped to the same shared ontology.
- We use an event extraction module that pairs any textual and structural property with ontological properties.

In the next sections, we will explain each element in more detail.

5.3 KYOTO Overview

The KYOTO system is designed to exploit rich semantic background knowledge packaged in many different linguistic expressions. Because background knowledge plays a major role, KYOTO allows communities to model terms and concepts in their domain, which helps to extract events from text. As such, KYOTO follows a knowledge-rich approach to interpret text that can be extended, tuned and maintained for specific domains. Nevertheless, the architecture of KYOTO is set up as a generic system that can model event structures in any text and any domain.

Figure 5.1 shows an overview of the process in which documents are processed through a pipeline of modules. The knowledge cycle starts with a set of source documents (at the left top side), which are converted to HTML format if necessary. Next, linguistic processors apply tokenization, segmentation, morpho-syntactic analysis and semantic processing to the text in different languages. In the current system, there are processors for English, Dutch, Italian, Spanish, Basque, Chinese and Japanese. The output of the linguistic processors is stored in an XML annotation format that is the same for all the languages, called the KYOTO Annotation Format (KAF, [4]). KAF incorporates proposals for standardized linguistic annotation of text and represents them in a layered structure, compatible with the Linguistic Annotation Framework (LAF, [11]). Once the text is represented in KAF, a series of semantic processing modules is launched that take KAF as input and produce KAF as output with a new conceptual interpration.The semantic processing involves the detection of multiword expressions, named-entities (persons, organizations, places, time-expressions), determining the most-likely synsets of words according to a given wordnet [6] and assigning ontological labels to textual units through the wordnet synsets. The result is that every element in the textual representation will get a corresponding semantic representation in terms of synsets and the associated ontological classes that apply to each synset. For the semantic processing of KAF, the system uses a knowledge base that contains wordnets in seven languages and a shared central ontology.

The KYOTO system then proceeds in two cycles (see Fig. 5.1). In the 1st cycle, we extract potentially relevant terms from the documents represented in KAF, such as *migratory fish* and *anadromous species*. Terms are normalized (sequences of) words that have sufficient frequency and/or many semantics relations with other terms in a set of documents for a domain. The terms are organized as a structured hierarchy and, wherever possible, related to existing concepts in the given knowledge base, i.e. wordnets for each language. For example in the case of

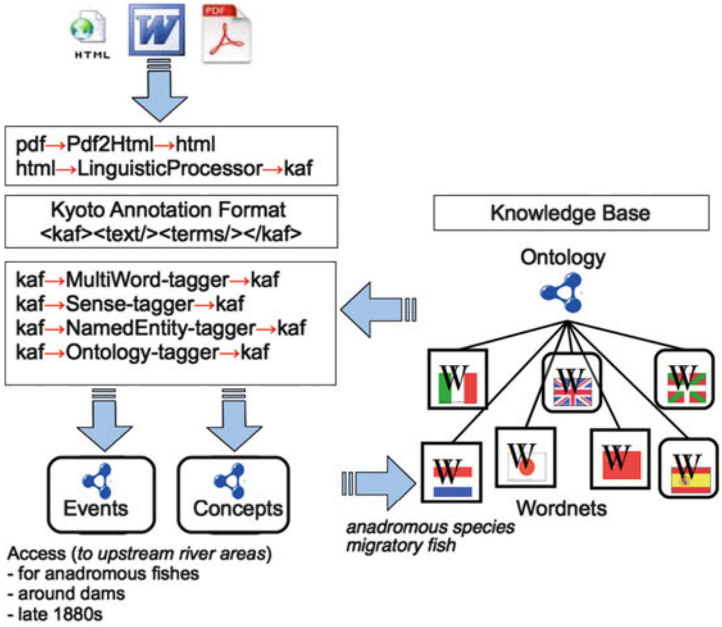


Fig. 5.1 Overview of the KYOTO architecture

migratory fish, the word-sense-disambiguation (WSD) module [1] will determine the most-likely sense of *fish* in the sequence and likewise determine the hypernym synset to which the new term will be connected. Since each wordnet is mapped to the central ontology, also the new terms are ultimately mapped to the ontology. The extended knowledge base is then used for processing new text, adding more precision to the interpretation: while *fish* and *migratory* have two meanings in the general WordNet, *migratory fish* will only have one in the extended WordNet. Customization and tuning of the processing can thus be done by adding more specific knowledge.

From the same KAF with semantic information, we also extract events in the 2nd cycle by so-called Kybots (Knowledge Yielding Robots). Kybots use a collection of profiles that represent patterns of information of interest. In the profile, conceptual relations are expressed using ontological and morpho-syntactic linguistic patterns, e.g. a noun with the ontology class *species* is followed by a verb with the class *change-of-location*. When a match is detected, the instantiation of the pattern is saved in a formal representation. Since the wordnets in different languages are mapped to the same ontology and the text in these languages is represented in the same KAF, similar patterns can easily be applied to multiple languages.

KAF plays an important role in the architecture of the system. In KAF, words, terms, constituents and syntactic dependencies are stored in separate layers with references across the structures. This makes it easier to harmonize the output of linguistic processors for different languages and to add new semantic layers to the

```
<KAF>
<text>
    <wf page="29" sent="770" wid="w10963">the</wf>
    <wf page="29" sent="770" wid="w10964">passage</wf>
    <wf page="29" sent="770" wid="w10965">of</wf>
    <wf page="29" sent="770" wid="w10966">migratory</wf>
    <wf page="29" sent="770" wid="w10967">fish</wf>
<text/>
<terms>
<term lemma="passage" pos="N" tid="t9032" type="open">
   <externalReferences>
   <externalRef conf="0.52" ref="eng-30-03895293-n" res="wneng3.0">
*        <externalRef ref="eng-30-00021939-n"  reftype="baseConcept" res="wn30g"/>
*        <externalRef ref="CommonSenseMapping.owl#geographical-object" reftype="sc_domainOf" res="ontology">
**            <externalRef reftype="SubClassOf" ref="CommonSenseMapping.owl#physical-place"/>
**            <externalRef reftype="SubClassOf" ref="ExtendedDnS.owl#non-agentive-physical-object"/>
**            <externalRef reftype="SubClassOf" ref="DOLCE-Lite.owl#physical-object"/>
**            <externalRef reftype="SubClassOf" ref="DOLCE-Lite.owl#physical-endurant"/>
**            <externalRef reftype="SubClassOf" ref="DOLCE-Lite.owl#endurant"/>
*        </externalRef>
*        <externalRef ref="Kyoto#connect" reftype="sc_participantOf" res="ontology"/>
*        <externalRef ref="Kyoto#has-path" reftype="sc_playRole" res="ontology"/>
    </externalRef>
   <externalRef conf="0.061" ref="eng-30-07310642-n" res="wneng3.0">
*        <externalRef ref="eng-30-07283608-n"  reftype="baseConcept" res="wn30g"/>
*        <externalRef ref="Kyoto#natural_event-eng-3.0-07283608-n" reftype="sc_subClassOf" res="ontology">
**            <externalRef reftype="SubClassOf" ref="DOLCE-Lite.owl#event"/>
**            <externalRef reftype="SubClassOf" ref="DOLCE-Lite.owl#perdurant"/>
*        </externalRef>
    </externalRef>
</externalReferences>
</term>
<!-- etc. -->
<term lemma="migratory fish" pos="N" tid="t9035mw" type="open">
<externalReferences>
<externalRef conf="0.409837" ref="dw-eng-30-343-n" res="wneng3.0">
*        <externalRef ref="eng-30-02512053-n"  reftype="baseConcept" res="wn30g"/>
*        <externalRef ref="Kyoto#fish-eng-3.0-02512053-n" reftype="sc_domainOf" res="ontology">
**            <externalRef reftype="SubClassOf" ref="Kyoto#animal-eng-3.0-00015388-n"/>
**            <!-- etc.. -->
**            <externalRef reftype="SubClassOf" ref="DOLCE-Lite.owl#physical-endurant"/>
*        </externalRef>
*        <externalRef ref="Kyoto#migration" reftype="sc_participantOf" res="ontology">
**            <externalRef reftype="SubClassOf" ref="Kyoto#active-change-of-location"/>
**            <externalRef reftype="Kyoto#done-by" ref="Collections.owl#physical-plurality"/>
**            <externalRef reftype="SubClassOf" ref="Kyoto#change_of_location-eng-3.0-00280586-n"/>
**            <externalRef reftype="Kyoto#has-source" ref="DOLCE-Lite.owl#particular"/>
**            <externalRef reftype="Kyoto#has-path" ref="DOLCE-Lite.owl#particular"/>
**            <externalRef reftype="Kyoto#has-destination" ref="DOLCE-Lite.owl#particular"/>
**            <externalRef reftype="SubClassOf" ref="Kyoto#change-eng-3.0-00191142-n"/>
**            <externalRef reftype="SubClassOf" ref="DOLCE-Lite.owl#accomplishment"/>
**            <externalRef reftype="SubClassOf" ref="DOLCE-Lite.owl#event"/>
**            <externalRef reftype="SubClassOf" ref="DOLCE-Lite.owl#perdurant"/>
*        </externalRef>
*        <externalRef ref="Kyoto#done-by" reftype="sc_playRole" res="ontology">
**            <externalRef reftype="InverseObjectProperties" ref="Kyoto#active-participant-in"/>
**            <externalRef reftype="SubObjectPropertyOf" ref="Kyoto#done-by"/>
**            <externalRef reftype="SubObjectPropertyOf"
**              ref="FunctionalParticipation.owl#functional-participant"/>
**            <externalRef reftype="SubObjectPropertyOf" ref="DOLCE-Lite.owl#participant"/>
*        </externalRef>
    </externalRef>
</externalReferences>
</term>
</terms>
<!-- Additional layers (chunking, dependencies, ...) -->
</KAF>
```

Fig. 5.2 Terms in KAF (in *blue*) expanded with ontological tags. Ontological classes from direct mappings are marked with '*' and implied ontological classes are marked with '**' and in *red*

basic output, when needed. All semantic modules for interpreting textual elements into conceptual structures draw their input from this structure. This means that these modules are the same for all the involved languages, resulting in further interoperability. Figure 5.2 shows in blue and without the prefix '*' a shortened example of a KAF structure, representing a text and a term layer. The text layer shows five sequential word tokens (*the passage of migratory fish*) and the term layer shows four corresponding *terms*. Terms have attributes such as lemma, part-of-speech and a unique identifier. Furthermore, they have elements (*span*) that refer back to the word tokens that make up the terms and references to external sources (*externalReferences*) which represent the semantic interpretation of the textual elements. In the case of *passage*, we see 2 out of a list of 10 wordnet synsets representing its different meanings, where the *conf* attribute indicates the score of the WSD [1]. External references can be nested and here we show the mappings to ontological classes for the first two senses only, prefixed with a '*'. The ontology and the mapping relations are explained in more detail below. In the case of the multiword *migratory fish*, we have a single term that refers back to two word tokens and there is only a single meaning, thanks to the acquisition of concepts in the 1st cycle, which led to the extension of wordnet with the concept *migratory fish*.

5.4 Ontological and Lexical Background Knowledge

Defining terms and concepts in a domain is an important step towards the disclosure of knowledge. In many cases, communities already have large quantities of (semi-) structured vocabularies and thesauri. Modeling these terms and concepts is a huge integration task, possibly involving millions of concepts and relations. To cope with these different types of knowledge, we designed a three-layered knowledge model [21] along the notion of the division of labor [20]. According to this model, we assume that domain experts know how to distinguish rigid and disjoint types of things (as defined by Guarino and Welty [8]) in their domain. There is no need to define the identity criteria for fishes such as *Alosa sapidissima* and *Brevoortia tyrannus* for computers. The simple fact that these are subclasses of an ontological type (e.g. fish) is sufficient to know that they are disjoint, each with a unique set of properties: *Alosa sapidissima* will never become *Brevoortia tyrannus*. Instead, it is more important to model the actual processes and states in which these rigid types of fish can be involved: e.g. being *invasive*, *endangered*. Specialists can consult encyclopedia or text books to find static knowledge about types of species but they urgently need to access textual sources to learn about new trends and environmental changes in local areas over time. We thus argue that software that supports such specialists needs to know what these processes and states are to mine informative events from text. Following these observations, we distinguished three knowledge layers:

1. Domain and background vocabularies in different languages
2. Wordnets in different languages
3. A central ontology shared by all languages

The first layer consists of large volumes of background knowledge and new terms learned from text collections in the domain. This layer is automatically linked to wordnets in different languages. All the wordnets are linked to the English WordNet. The wordnets represent the 2nd layer of knowledge, which is linked to the 3rd layer: the central ontology. Each of these layers has an internal semantic structure, connecting specific concepts to more general concepts and it has specific mapping relations to the next layer. In this model, it is not necessary to have a mapping relation between all the concepts across the resources, since we can use the internal relations in each resource to find a more general concept with a mapping. Whenever we come across a term such as *Ethmidium maculatum* which is not in WordNet, we traverse the relations in a species database[2] until we find a more general concept (*Brevoortia*) that is matched to WordNet. Next, we traverse the hypernym relations in WordNet until we find a synset (*fish genus*) that is matched to the ontology. When combining vocabularies, we assume the principle that all concepts related to more general concepts are rigid-subtypes unless there is evidence to the contrary. Consequently, we need a specification for non-rigid terms, such as *alien invasive fish* and *migratory fish* to explain (1) that they are **not** rigid types of fish and (2) what their role is in vital processes and conditions. In the next sections, we describe the ontology and the formal model for these relations in more detail.

5.4.1 Ontology

The ontology consists of around 2,000 classes divided over three layers [9]. The top layer is based on DOLCE[3] [15] and OntoWordNet [7]. The second layer are the Base Concepts[4] which cover an intermediate level of abstraction for all nominal and verbal WordNet synsets [12]. Base concepts are hypernym synsets that have relatively many relations to other synsets and cover all different branches of the wordnet hierarchy. Examples of Base Concepts are: *building, vehicle, animal, plant, change, move, size, weight*. They provide an interface from the ontology to a complete wordnet. A third layer consists of domain classes introduced for detecting events and qualities in a particular domain (i.e. environment).

A mapping for every synset in the English WordNet is provided to the ontology, where the so-called Base Concepts guarantee that there is such a mapping through the hyponymy relations: 114,016 mappings to the Base Concepts, 185,666 mappings to the central ontology together with 30,000 mappings from ontology labels to implications in the ontology.[5] The word-to-concept mapping also harmonize predicate information across different parts-of-speech. For instance, *migratory events* are represented by different synsets such as the verb *migrate*, the noun *migration* and the adjective *migratory*, which all inherit the same ontological information corresponding to the *active-change-of-location* class. Furthermore, through the equivalence relations of wordnets in other languages to the English WordNet, this semantic framework can also be applied to other languages.

[2] http://www.sp2000.org/

[3] DOLCE-Lite-Plus version 3.9.7

[4] http://adimen.si.ehu.es/web/BLC

[5] This knowledge model is freely available through the KYOTO website as open-source data.

Table 5.1 Rigid and non-rigid synset to ontology mappings

wn:allis shad	hypernym	wn:shad
wn:shad	hypernym	wn:fish
wn:fish	sc_equivalenceOf	ont:fish
wn: anadromous fish	hypernym	wn:migratory fish
wn:migratory fish	hypernym	wn:fish
	sc_domainOf	ont:fish
	sc_playRole	ont:done-by
	sc_participantOf	ont:migration
wn:fish migration	sc_subcassOf	ont:migration (perdurant)
	sc_hasParticipant	ont:fish
	sc_hasRole	ont:done-by
wn:air pollution	sc_subcassOf	ont:pollution (perdurant)
	sc_hasParticipant	ont:air
	sc_hasRole	ont:patient
wn:nitrogen pollution	sc_subcassOf	ont:pollution (perdurant)
	sc_hasParticipant	ont:nitrogen
	sc_hasRole	ont:done-by

5.4.2 Wordnet to Ontology Mappings

Relations from wordnet synsets to the ontology are used to differentiate between rigid and non-rigid concepts. This is done in the following way, where the prefix sc_ stands for synset-to-class:

sc_equivalenceOf: the synset is fully equivalent to the ontological class and inherits all properties; the synset is Rigid;

sc_ subclassOf: the synset is a proper subclass of the ontological class and inherits all properties; the synset is Rigid;

sc_domainOf: the synset is not a proper subclass of the ontological class and is not disjoint (therefore orthogonal) with other synsets that are mapped to the same class; the synset is therefore non-Rigid but still inherits all properties of the target ontology class; the synset is also related to a Role with a sc_playRole relation;

sc_playRole: the synset denotes instances for which the context of the Role applies for some period of time but this is not essential for the existence of the instances, i.e. if the context ceases to exist then the instances may still exist [16];

sc_participantOf: instances of the concept (denoted by the synset) participate in some perdurant class of the ontology, where the specific role relation is indicated by a sc_playRole mapping;

Table 5.1 shows some examples. Using these relations, we can express that the synset *alis shad* is a proper subclassOf the ontological type *fish* because it is related to the synset *shad* as a hypernym, which is related to the syset *fish* as a hypernym, where the latter has an sc_equivalenceOf mapping with the ontological type. For newly acquired non-rigid concepts, such as *anadromous fish* and *migratory fish*, we create internal wordnet hypernym relations but also an explicit mapping to the ontology to indicate their non-rigid status. This mapping indicates that the synset for *migratory fish* is used to refer to instances of *fish* (not subclasses!), where the domain is restricted to *fish*. Furthermore, these instances participate in the process of migration in the role of done-by. The fact that *anadromous fish* is a hyponym of *migratory fish* implies that it is also non-rigid by definition, whereas the fact that *migratory fish* is a hyponym of *fish* does not imply that the former is rigid. Rigidity is not transitive along hypernym relations but non-rigidity is. The properties of the process migration are further defined in the ontology. As a subclass of *active-change-of-location*, it involves an endurant as a *done-by* participant and it has further roles *has-source*, *has-path* and *has-destination*.[6]

Ideally, all processes and states that can be applied to endurants should be defined in the ontology. This may hold for most verbs and adjectives in languages, which do not tend to extend in specific domains and are part of the general vocabulary. However, domain specific text contain many new nominal terms that refer to domain-specific processes and states, e.g. *fish migration*, *air pollution* or *nitrogen pollution*. These terms are equally relevant as their counter-parts that refer to endurants involved in similar processes, e.g. *migratory fish*, *polluted air*, *polluting nitrogen*. As shown in Table 5.1, we therefore use the reverse participant and role mappings to define such processes as subclasses of more general processes involving specific participants in a specified role.

Our model extends other existing WordNet to ontology mappings. For instance in the SUMO to Wordnet mapping [17], only sc_equivalenceOf and sc_subclassOf relations are used, represented by the symbols = and + respectively. The SUMO-Wordnet mapping likewise does not systematically distinguish rigid from non-rigid synsets. Through the mapping relations, we keep the ontology relatively small and compact whereas we can still define the richness of the vocabularies of languages in a precise way. To summarize, event relations can be derived in the following ways in KYOTO:

1. Wordnet relations between synsets that express role relations between events and participants. These are still rare in the English WordNet.
2. Wordnet to ontology mappings from event synsets to ontological participants and from participant synsets to ontological events
3. Ontological axioms that express role relations between events and participants
4. Inheritance in Wordnet of relations through hyponymy relations and in the ontology through subclass relations

In the next sections, we will explain how we exploit these options for inserting the semantic information in the KAF representations and to use these for extracting events and event relations in texts.

[6] The mapping relations from wordnet to the ontology, need to satisfy the constraints of the ontology, i.e. only roles can be expressed that are compatible with the role-schema of the process in which they participate.

5.5 Off-Line Reasoning and Ontological Tagging

The ontological tagging represents the last phase in the KYOTO annotation pipeline described in Sect. 5.2. It consists of a three-step module to enrich the KAF documents with knowledge derived from the ontology. For each synset connected to a term, we first add the Base Concepts to which the synset is related through the wordnet hypernym relations. Next, through the synset to ontology mapping, we add the corresponding ontology type with appropriate relations. Once each synset is annotated with its ontology type, we finally insert the full set of ontological implications that follow from the ontology. The ontological implications are extracted from the OWL representation of the ontology and stored in a static table for all ontological classes. The main purpose is to optimize the performance of the mining module over large quantities of documents, but it is also very useful for debugging.

Figure 5.2 shows, in red and prefixed with '*' and '**', a fragment of the result of onto-tagging for the correct meanings of *passage* and *migratory fish*. Compared to the blue parts we see an additional reference to a Base Concept and the ontological mappings have been expanded with a series of implications (marked with '**') resulting from the offline reasoning. For example, the implications reflect the subclass hierarchy of the ontology and indicate that the first sense of *passage* is an *endurant* and the second sense is a *perdurant*. In the case of *migratory fish*, we see that the mapping as a participant to the ontology class *Kyoto#migration* gives us the implied information that this event also involves the roles *has-source*, *has-destination* and *has-path*.

There are a number of advantages for expanding the KAF representation with ontological implications. First of all, we can now formulate patterns of ontological classes or base concepts instead of looking for sequences of words or synsets. We thus need less patterns to capture more event structures. It is relatively easy to experiment with patterns at different levels of specificity to find the optimal balance between precision and recall (e.g. searching either for *perdurants*, *accomplishment* or *changes of locations*). Secondly by making the implicit ontological statements explicit, we can find the same relations in many different expressions with different surface realizations: *fish migration*, *migratory fish*, *migration of fish*, *fishes that migrate* etc. Since these expressions share the same ontological implications, we can apply similar patterns for the extraction of events. Thirdly, event-participant relations that are not overtly expressed but are semantically implied are still available for matching and can be used to create relations with surrounding expressions, e.g. *passage* can fill the has-path role of *Kyoto#migration* that is implied by *migratory fish*. The same implication will also be represented for terms such as *anadromous fish* as a hypernym of *migratory fish*. Furthermore, the implications will be represented in the same way across different languages, thus facilitating cross-lingual extraction of events. Finally, onto-tagging is a kind of off-line ontological reasoning through which the pattern matching can be relatively easy, fast and robust. There is one big disadvantage to this approach in that the size of the KAF files is expanded by a factor of 20.

5.6 Event Extraction

Kybots (Knowledge Yielding Robots) are programs that find sequences of concepts to extract instances of events, participants and relations in KAF documents. The Kybot server loads a set of profiles that express patterns of such sequences and compiles them into Kybots that scan enriched documents in KAF for matches. In case of a match, the Kybot server will output elements from the text into a a specified output format. Due to our ontology insertion method, these KAF files include all possible implications of all word meanings of the text, which can all be used for matching in the profiles. The Kybot module uses two different methods to find event-participant relations:

1. Profiles that represent sequences of terms exhibiting event-participant relations
2. Complex terms that exhibit an event-participant relation as part of their meaning

The Kybot profiles have a declarative XML format, which describes general morpho-syntactic patterns and semantic conditions on sequences of terms. Linguistic patterns can include morphological and lexical constraints but also semantic conditions that must hold for terms. Kybot are thus able to search for term lemmas or part-of-speech tags but also for terms linked to ontological process and states using the mappings described in before. Figure 5.3 presents an example of a profile. The profiles consist of three main parts:

- Variable declaration (`<variables>` element): defines the search entities e.g.: **x** (denoting terms whose part-of-speech is noun and lemma is not *people*), **y** (which are terms whose lemma is *move*, *migrate* or *travel*), **p** (which are the prepositions *into* or *to*) and **z** (terms linked through one of its synsets to a subclass of the ontological class *CommonSenseMapping.owl#geographical-object*).
- Relations among variables (`<rel>` element): specifies the relations among the previously defined variables e.g.: **y** is the main pivot, variable **x** must precede variable **y** in the same sentence, variable **p** follows **y** and variable **z** must follow variable **p**. Thus, this relation declares patterns like '**x** → **y** → **p** → **z**' in a sentence.
- Output template: describes the output to be produced for every matching structure e.g.: each match generates a new event targeting term **y**, which becomes the main term of the event with two roles: the 'done-by' role filled by term **x** and 'destination-of' role, filled by **z**.

The profile in Fig. 5.3 would match a sentence such as *Menhaden migrate into Chesapeake Bay* and output the structure of Fig. 5.4. This example shows that we can directly use any ontological class that is inserted in KAF to constraint the variables. Likewise, we can formulate patterns that capture any ontological feature that is either directly or indirectly associated with a word meaning in the text, to express either an event, a participant or a relation. We can therefore replace lexical constraints such as the disjunction *move or migrate or travel* by a more

```
<kprofile>
 <variables>
  <var name="x" type="term" pos="N" lemma="! people"/>
  <var name="y" type="term"
       lemma="move | migrate | travel"/>
  <var name="p" type="term"  pos="P" lemma="into | to"/>
  <var name="z" type="term"
       ref="CommonSenseMapping.owl#geographical-object"
       reftype="SubClassOf"/>
 </variables>
 <relations>
  <root span="y"/>
  <rel span="x" pivot="y" direction="preceding"/>
  <rel span="p" pivot="y" direction="following"/>
  <rel span="z" pivot="p" direction="following"/>
 </relations>
 <events>
  <event target="$y/@tid" lemma="$y/@lemma" pos="$y/@pos"/>
  <role target="$x/@tid" rtype="done-by" lemma="$x/@lemma"/>
  <role target="$z/@tid" rtype="destination-of" lemma="$z/@lemma"/>
 </events>
</kprofile>
```

Fig. 5.3 Example of a Kybot profile

```
<event eid="e97" target="t9643" lemma="migrate" pos="V"
  synset="eng-30-01857093-v" rank="0.5"/>
<role rid="r191" event="e97" target="t9646mw"
  lemma="chesapeake bay" pos="N" rtype="destination-of"
     synset="eng-30-09243405-n" rank="1.0""/>
<role rid="r84" event="e97" lemma="menhaden"
    target="t9642" rtype="Kyoto#done-by"/>
```

Fig. 5.4 Output structure resulting from a Kybot profile

powerful ontological constraint such as the class *Kyoto#active-change-of-location*. Similarly, we can replace the exclusion of the lemma *people* by the ontology class *Kyoto#person-eng-3.0-00007846-n*, which captures all words and expressions in wordnet that relate to this class. The resulting profile would not only match many more expressions in English but, after adapting the prepositions, would also work for many other languages linked to the same ontology through their wordnet. Through closure of the ontology and wordnets by the Base Concepts, i.e. every synset in wordnet is linked to a Base Concept and every Base Concept is mapped to the ontology, we can thus guarantee maximal coverage of the profiles. It is therefore possible to detect similar event information within and across documents even if expressed differently and in different languages.

One drawback of the profiles is that they can only relate sequences of distinct terms that represent events and participants. In many cases, the event and a participant are both implied by a single term. For example, role-denoting terms, such as *migrant*, *prey*, *predator*, refer to participants and implicitly also to the event in which they are involved. Similarly, event-denoting terms such as *migration* already imply participants even when they are not explicitly mentioned in the surroundings of the term. Actually, one of the effects of acquiring terms for a specific domain

```
<event eid="e3"
       lemma="Kyoto\#change\_of\_location-eng-3.0-00280586-n"
       target="t8570mw" profile_id="complex_term"/>
<role rid="r3" event="e3" lemma="migratory fish"
       target="t8570mw" rtype="Kyoto\#done-by"
       profile_id="complex\_term"/>

<event eid="e28" lemma="spawn"
     target="t8575" profile_id="complex_term"/>
 <role rid="r42" event="e28"
       lemma="Kyoto\#fish-eng-3.0-02512053-n"
       target="t8575" rtype="Kyoto\#done-by"
       profile_id="complex\_term"/>
```

Fig. 5.5 Events and participants extracted from the terms *migratory fish* and *spawn*

is that many multiword expressions, such as *migratory fish*, *murky water* and *crab exploitation*, become single terms in our KAF representation and likewise cannot be matched through the Kybot profiles. Whereas the domain acquisition adds semantics and precision for these words, we loose the possibility to detect the sequence of elements. To be able to still exploit the semantic richness of such terms (both generic and domain specific), we defined special kybots which extract event-participant relations that are implicit. The so-called complex-term process works in two ways:

1. Search for terms that are events (subclasses of perdurant) and look for any role that is defined within the same set of ontological implications related to the same word meaning;
2. Search for terms that are potential participants (endurants) and look for roles and events expressed within the same set of ontological implications related to the same word meaning;

In the first case, the Kybots will output an event represented by the term and a role by the ontological class of the role that is defined. In the second case, the Kybots output an event represented by the ontological class whereas the participant is represented by the term. Figure 5.5 shows the event representation for two such terms *migratory fish* and *spawn*. In the case of the domain term *migratory fish*, the term is the lemma for the role *done-by* and the ontological class *Kyoto#change_of_location-eng-3.0-00280586-n* is given as the lemma for the event. Both the role and the event have the same term identifier as the target. In the case of the generic verb *spawn*, we see that the verb is the lemma for the event and that the ontological class "Kyoto#fish-eng-3.0-02512053-n" is the lemma for the role. Again, both the event and the role have the same target term identifier.

The representation of the implied event and the implied role is important because they do not only capture relations outside the scope of the profiles but can also connect to other elements in the text. In the surrounding of *migratory fish*, we may find concepts for the source, path or destination of the *migration*. In the surroundings of the verb *spawn*, we can expect other concepts related to *fish* even when these *fish* are not explicitly mentioned.

5.7 Experimental Results

To evaluate the platform, we carried out an in-depth evaluation on a single document and we applied the system to a large volume of documents. Finally, we applied the same system to another domain (medical) and to another language (Dutch).

5.7.1 In-Depth Evaluation

The event structure in KYOTO is rather specific and events can be complex. To be able to compare our results with other systems and gold-standards, we defined a more neutral triplet format

```
<R, E, P>
```

where R is a relation, E is a set of word tokens representing the event and P is a set of word tokens representing a participant. If an event has multiple participants, a separate triplet is created for each event-participant pair. The triplet identifier is used to mark which triplets relate to the same event. The Kybot output shown in Fig. 5.4 is then converted to the following two triplets, where the target term identifiers are converted to word token identifiers:

```
<triplet id="941" profile_id="" relation="destination-of">
    <eventids comment="migrate">
        <event id="w11698"/>
    </eventids>
    <participantids comment="Chesapeake Bay">
        <participant id="w11700"/><participant id="w11701"/>
    </participantids>
</triplet>
<triplet id="941" profile_id="" relation="done-by">
    <eventids comment="migrate">
        <event id="w11698"/>
    </eventids>
    <participantids comment="Menhaden">
        <participant id="w11697"/>
    </participantids>
</triplet>
```

A range of (possibly disjoint) token identifiers can be given in a triplet, as shown for *Chesapeake Bay*. Events and participants across triplets therefore match if at least one identifier overlaps, while the relation is the same. Abundance of identifiers is blocked. **P**recision, **R**ecall and **F**-measure are then calculated as follows, where C is the correct system triplets, N_{GS} is the total number of gold standard triplets and N_{S} is the total number of system triplets:

$$P = \frac{C}{N_{\mathrm{S}}} \qquad R = \frac{C}{N_{\mathrm{GS}}} \qquad F = \frac{2(PR)}{(P + R)}$$

Table 5.2 Document statistics

	Nouns	Verbs	Adjectives
Nr. of Terms	893	375	201
Sense tokens:	3,013	3,668	680
Average polysemy	3	10	3
Sense types:	1,065	1,007	353
Base concept tokens:	3,013	3,668	680
Base concept types:	144	223	75
Ontology tokens	14,530	24,763	2,717
Ontology types	573	484	160
Implied ontology tokens:	73,639	126,275	10,262
Implied ontology types:	524	480	214

Table 5.3 Synset to ontology mappings in the text

Mapping	Noun	Verb	Mapping	Noun	Verb
sc_domainOf	63		sc_resultOf	268	30
sc_hasParticipant	294	1,486	sc_simpleCauseOf	43	179
sc_participantOf	686	14	sc_hasCoParticipant	26	
sc_hasRole	251	1,154	sc_playCoRole	26	
sc_playRole	402	6	sc_equivalentOf	463	1,634
sc_subClassOf	3,978		Total	7,123	4,613

For the in-depth evaluation, we took a single document[7] about the Chesapeake Bay, a large estuary in the US. The document has 16,145 word tokens. We manually annotated 132 sentences (2,927 word tokens) from the document with events, participants and their roles. This resulted in 263 events and 470 triplets. We processed the text using the KYOTO system, where we used the generic English WordNet, the KYOTO ontology and a domain wordnet with 990 terms from the environment that have been mapped to the generic WordNet and to the ontology (including the term *migratory fish*). Table 5.2 shows some of the statistics for the document after processing it with the KYOTO system. Average polysemy for nouns and adjectives is three but ten for verbs. Consequently, almost three times as many nouns in the text yield the same number sense meanings and a similar amount of base concepts as the verbs. Furthermore, we see that the nouns result in 14 K mappings to ontology classes, the verbs in 24 K mappings and the adjectives in 2 K mappings. Even though verbs map to many more classes, in the end this boils down to the same proportion of distinct classes (about 500 different types, which is 25 % of the ontology). The ontology classes yield more implied ontology classes, which are classes resulting from the semantics in the ontology. For the verbs 126 K classes apply, which is 1.7 times the amount of classes that apply to the nouns. Table 5.3 shows the important synset-to-ontology mappings for events

[7] www.acb-online.org/pubs/BayBarometer2008Web.pdf

Table 5.4 Baseline and Kybot results on a gold standard of 132 sentences with 263 events and 470 triplets

	Baseline	Kprofiles	Cterms	Profiles-Cterms	Profiles-Wsd	Profiles-Wsd-Cterms
Nr. events	1762	773	32	795	719	741
Nr. correct	319	239	16	249	227	237
Precision	0.18	0.31	0.50	0.31	0.32	0.32
Recall	1.0	0.91	0.06	0.95	0.86	0.90
F-measure	0.32	0.46	0.11	0.47	0.46	0.47
Nr. triplets	4,688	644	19	663	511	530
Nr. correct	131	181	10	191	164	174
Precision	0.03	0.28	0.53	0.29	0.32	0.33
Recall	0.28	0.39	0.02	0.41	0.35	0.37
F-measure	0.05	0.32	0.04	0.34	0.33	0.35

and participants. Obviously, *sc_equivalentOf* and *sc_subClassOf* are most frequent (62 %) but the remainder mostly introduces event-role relations.

To evaluate our system, we created 261 profiles that were applied to the 132 sentences of the gold standard. The profiles consider all ontological classes associated with all meanings. However, these meanings were scored by the WSD system. We therefore considered two variants of the system: one considering all meanings and one considering only the meanings with the highest rank if there was a choice to be made between alternative interpretations of profiles (see [22] for more details on the role of WSD in the process of event extraction). In addition to the profiles, we also extracted relations through the complex-term approach. Combining these options, we get the following variants:

1. Kprofiles: applying the 261 profiles to all different meanings of words
2. Cterms: detecting event-participant relations implied by the meaning of a single term (possibly a multiword term)
3. Profiles-Cterms: combining the results of 1 and 2
4. Kprofiles-Wsd: applying profiles only to the meanings with the highest word-sense-disambiguation score if there is a choice between profiles
5. Profiles-Wsd-Cterms: combining 4 with 2

As a baseline, we created triplets for all heads of constituents in a single sentence according to the constituent representation of the text in KAF. The baseline generates 4,688 triplets for the annotated sentences. Since there is no relation predicted, we assume the most-frequent patient relation for all.

Table 5.4 shows the results of the baseline and the Kybot variants. The top part of the table shows the results for detecting the 263 gold standard events. The baseline and Kybot profiles have high recall (100 and 91 %). The baseline gives an extremely low precision, whereas the precision of the Kybot profiles is 31 %. Precision gets slightly higher when we apply WSD. The Cterms heuristics has low recall but higher precision (50 %). We get the best f-measure by combining profiles, WSD and Cterms. For the triplets in the lower part of the table, we see similar results even

Table 5.5 Baseline and Kybot results on four sentences containing *migration*, *migratory* and *migrate*, representing 20 events and 43 triplets in the gold-standard

	Baseline	Kprofiles	Cterms	Profiles-Cterms	Profiles-Wsd	Profiles-Wsd-Cterms
Nr. events	79	48	3	48	42	42
Nr. correct	20	17	3	17	15	15
Precision	0.25	0.35	1.00	0.35	0.36	0.36
Recall	1.00	0.85	0.15	0.85	0.75	0.75
F-measure	0.40	0.50	0.26	0.50	0.48	0.48
Nr. triplets	344	52	6	56	42	46
Nr. correct	5	8	3	10	8	10
Precision	0.01	0.15	0.50	0.18	0.19	0.22
Recall	0.12	0.19	0.07	0.23	0.19	0.23
F-measure	0.03	0.17	0.12	0.20	0.19	0.22

though the task is more difficult. Again, the Cterms approach has highest precision (53 %) and lowest recall and WSD adds precision to the profiles. We obtain the highest f-measure by combining them.

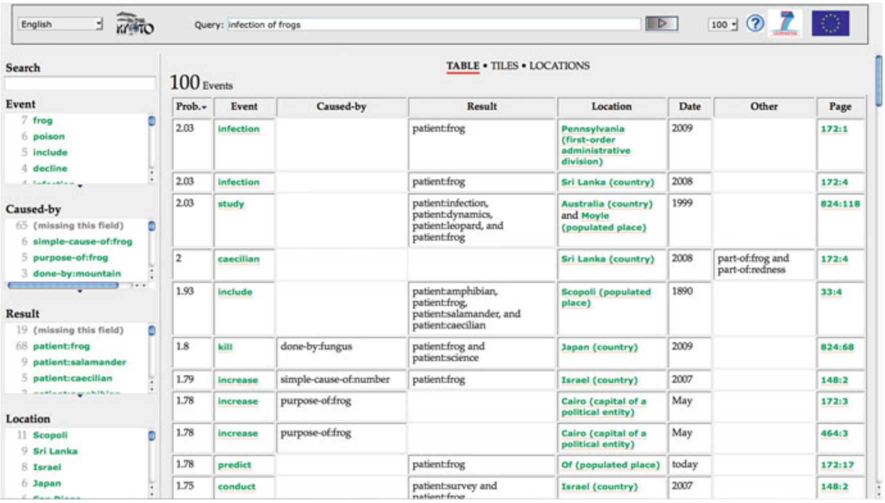
We also measured the results for four sentences (including examples 2, 4 and 7 from Sect. 5.2) that explicitly refer to migration using nominal, adjectival and verbal forms. The results are shown in Table 5.5. We see that the profiles perform slightly better on events but much worse on the triplets. The Cterms, on the other hand, perform much better on both events and triplets. Through the Cterms, the combined results recover a little but the best results still have a lower f-measure of 22 % compared to 35 %. This shows that the examples we considered are more complex than average compared to the gold-standard. It also shows that the Cterm approach can significantly contribute to the precision and recall of the system if sufficient terms are added to the knowledge base. In the current system, we added only 990 terms for the domain.

5.7.2 *Large Scale Evaluation*

The 291 English profiles have been optimized to extract the relations from a single document on environmental issues. After that they have been applied to almost 9,000 documents on environmental issue from various sources. We also applied the same profiles to another domain without adaptation: seven documents on medical breast cancer. These documents together contain about 25 million words and the profiles extracted 890 thousand events. Table 5.6 gives overview of the extracted data.

Table 5.6 Events extracted during the KYOTO project

	Docs	Word tokens	Events	Roles	Date refs	Dates	Places	Countries
Estuary	4,625	3,091,842	470,762	231,630	102,653	1,168	2,409	176
WWF International	1,174	1,966,914	264,743	331,391	38,057	711	1,224	146
Journal of Environmental Biology	791	3,440,611	23,406	27,782	51,188	696	2,306	82
European Environment Agency	713	4,814,647	47,355	58,628	105,952	662	1,348	93
Hydrology and Earth System Sciences	1,355	11,228,175	71,781	85,276	157,057	2,380	4,407	116
Medical breast cancer protocols	7	110,501	8,416	15,984				
Total	8,758	24,695,387	890,558	757,553	463,025			



100 Events

Prob.	Event	Caused-by	Result	Location	Date	Other	Page
2.03	infection		patient:frog	Pennsylvania (first-order administrative division)	2009		172:1
2.03	infection		patient:frog	Sri Lanka (country)	2008		172:4
2.03	study		patient:infection, patient:dynamics, patient:leopard, and patient:frog	Australia (country) and Moyle (populated place)	1999		824:118
2	caecilian			Sri Lanka (country)	2008	part-of:frog and part-of:redness	172:4
1.93	include		patient:amphibian, patient:frog, patient:salamander, and patient:caecilian	Scopoli (populated place)	1890		33:4
1.8	kill	done-by:fungus	patient:frog and patient:science	Japan (country)	2009		824:68
1.79	increase	simple-cause-of:number	patient:frog	Israel (country)	2007		148:2
1.78	increase	purpose-of:frog		Cairo (capital of a political entity)	May		172:3
1.78	increase	purpose-of:frog		Cairo (capital of a political entity)	May		464:3
1.78	predict		patient:frog	Of (populated place)	today		172:17
1.75	conduct		patient:survey and patient:frog	Israel (country)	2007		148:2

Fig. 5.6 Search results in table form for the query *infection of frogs*

All the data and the Kybot profiles can be downloaded from the KYOTO website. They are available in the KYOTO output format and in RDF format. Since most events are placed in time and space, we can consider them as (partial) descriptions of facts. As suggested above, we can compare events within the same region and time-frame. To illustrate this, we developed a semantic search system that uses a multi-lingual index on Kybot facts. We index the facts by the lemmas, the synset-ids and the synset-ids of the hypernyms for each event and role.[8]

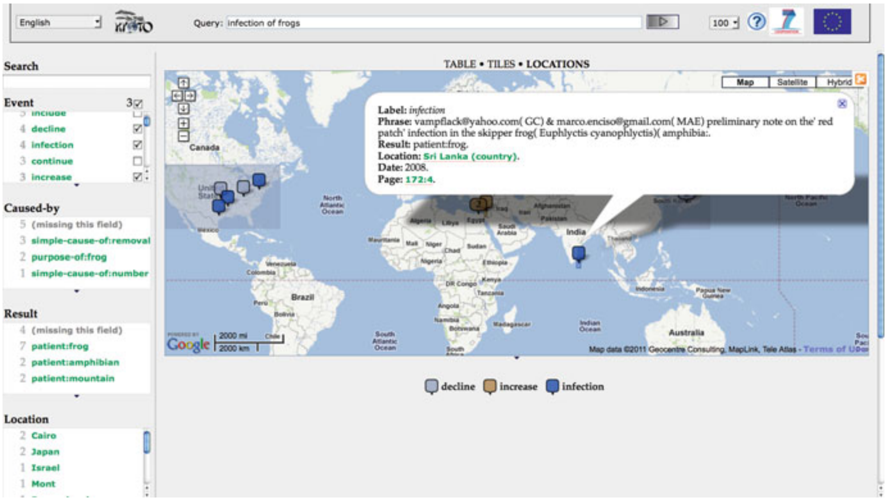
To search the indexed facts, a client program was developed through the Exhibit API.[9] Exhibit [10] consists of Java-script packages that provide advanced functionality to display structured data. The structured data can be published by any server (e.g. as Google spreadsheets) and are loaded in the browser of the user together with the Java-script. The local database of the user is accessed to further present the data. For KYOTO, the retrieved data are converted to a Json structure.

Queries are first lemmatized and sent to a word-sense-disambiguation server to obtain the most likely concepts associated with the words. The client receives the facts that have been indexed, orders them by the strength of the matches, and displays the 100 best facts. The databases mentioned in Table 5.6 can be searched through the demo that is available on the Kyoto website.[10] Figure 5.6 shows a

[8] For cross-lingual retrieval, the lemmas have been translated to all the other languages in KYOTO, using the equivalences in the wordnets. The databases can be searched in any of the languages and the results are rendered in the query languages, regardless of the source language of the information.

[9] http://simile-widgets.org/wiki/Exhibit

[10] Follow the next URL to search in the Estuary database. Login with any name and any password: http://kyoto.irion.nl/kyoto/web/init.do?project=estuary_en&database=2&queryLg=en&query=



Search

Event 3☑
5 include □
4 decline ☑
4 infection ☑
3 continue □
3 increase ☑

Caused-by
5 (missing this field)
3 simple-cause-of:removal
2 purpose-of:frog
1 simple-cause-of:number

Result
4 (missing this field)
7 patient:frog
2 patient:amphibian
2 patient:mountain

Location
2 Cairo
2 Japan
1 Israel
1 Mont

Label: *infection*
Phrase: vampflack@yahoo.com(GC) & marco.enciso@gmail.com(MAE) preliminary note on the' red patch' infection in the skipper frog(Euphlyctis cyanophlyctis)(amphibia:.
Result: patient:frog.
Location: Sri Lanka (country).
Date: 2008.
Page: 172:4.

Fig. 5.7 Search results on Google map for filtered results of the query *infection of frogs*

screen dump of the results when searching for *infection of frogs* in the Estuary database. The results are shown in a structured table with columns for the probability (matching score), the lemma for the event, causal roles, result roles, the location, date, other roles and finally a list of pointers to the sources. The table can be sorted by each column and you can click on the cell values to obtain further details. At the left side, filter tables are given for each column. They list the unique values with their frequency. By clicking on these values, a selection from the full table is shown.

The Exhibit API lets you display the Json structure in different ways, among which on Google maps. This is shown in Fig. 5.7, where the results have been filtered by selecting the events *infection*, *decline* and *increase*. The events are depicted on the map and by clicking an event information from the source is shown as for the event located in Sri Lanka.

We carried out a user-evaluation on three different retrieval systems:

1. A standard text search with a Google-like result list;
2. A mashup system that converts the results from the standard text search into similar Exhibit tables;
3. The semantic search on the Kybot output;

Sixteen students and six environment professionals participated in the study. The participants had to answer six questions per system, after a short introduction and practice with each system. Different groups answered different questions with different systems and in different order. Across the different system, we could not measure any significant difference in the quality of the answers and the time to find the answers. We also asked the users to provide feedback through the SUS-tool (a tool that measures usability; [5]). The feedback showed that out of 20 subjects, most users preferred the benchmark tool over the semantic search. The standard text search system scored best on usability and learnability, probably because it matches the experience all users have with Google. The system acts in the way they expect, matching phrases and presenting the results with snippets. These same users are confused by the semantic search which finds matches through concepts rather than phrases. However, another (smaller) group of subjects disliked the benchmark because it did not enable them to refine their search term or search very effectively. They preferred the semantic search because of its extra functionality.

We believe that semantic search is disliked by *conservative* users, who wish to be able to use a tool immediately, and who prefer the presentation to be familiar, so that they do not have to spend time learning to use the tool. However, it is liked by more *adventurous* users, who will invest time to investigate the extra functionality if they believe it will help them to search more effectively in the end, and to find better information. Presumably, there is also a *middle* group which could be persuaded to adopt the tool if its user-friendliness were improved, and/or if they were shown its potential and how to use it by fellow workers.

5.7.3 Transferring to Another Language

An important aspect of the KYOTO system is the sharing of the central ontology and the possibility to extract semantic relations in different languages in a uniform way. To test the feasibility of sharing the same semantic backbone and transferring Kybot profiles, we carried out a transfer experiment from English to Dutch. We collected 93 Dutch documents on a Dutch estuary (the *Westerschelde*) and related topics. We created KAF files using the Dutch parser Alpino[11] and applied WSD to these KAF files using the Dutch wordnet.

To apply the profiles to the Dutch KAF documents, we need to apply the ontology tagger to the Dutch KAF. However, the tables map the English WordNet to the ontology and not the Dutch wordnet. We therefore generated Dutch variants of the tables on the basis of the equivalence relations between the Dutch wordnet and the English wordnet. For each Dutch synset, we looked up all the equivalent synsets in English, next we looked up the English synset in the ontology tag tables. If there was a match, we created an entry for the Dutch synset in the new table with the same mapping. Likewise, we created tables that match every Dutch synset to the English Base Concepts and to the ontology. Some Dutch synsets have no equivalence and some have multiple equivalences. We generated 145,189 Dutch synset to English Base Concept mappings (for comparison for English we have 114,477 mappings) and 326,667 Dutch synset to ontology mappings (186,383 for English). These ontology tag tables were used to insert the ontological implications into the Dutch KAF files.

[11] http://www.let.rug.nl/vannoord/alp/Alpino/

Table 5.7 Roles related to the noun *toename* (increase) and the verb *stijgen* (increase)

toename (increase)			*stijgen* (to increase)		
Lemma	Role	Freq.	Lemma	Role	Freq.
Aantal (number)	Patient	1	Bodem (ground)	Patient	1
Activiteit (activity)	Patient	1	Zeespiegel (sea level)	Patient	3
Consumptie (consumption)	Patient	16	Zeespiegel (sea level)	Done-by	1
Vervuiling (pollution)	Patient	16	Aarde (earth)	Simple-cause-of	4
Introductie (introduction)	Done-by	16	Aarde (earth)	Patient	4
Atmosfeer (atmosphere)	Generic-location	2			
Handel (trade)	Patient	4			
Druk (pressure)	Patient	4			

Finally, we adapted the 261 English Kybot profiles to replace all English specific elements by Dutch. This mainly involved:

- Replacing English prepositions and relative clause complementizers by Dutch
- Adapting the word order sequences for relative clauses in Dutch
- Adapting profiles including adverbials
- Eliminating profiles for multiword compounds which hardly occur in Dutch
- Eliminating profiles for explicit English structures that express causal relations

We kept all the ontological constraints exactly as they were for English. Only superficial syntactic properties were thus changed. It took us half-a-day to adapt the profiles for Dutch. From the original 261 English profiles, we obtained 134 Dutch profiles. We ran the profiles on the 93 Dutch KAF files (42,697 word tokens) and 65 profiles generated output. In terms of relations, we see a similar distribution as for English: the patient relation is most frequent, followed by relations such as generic-location, has-state and done-by. We did a preliminary inspection and the results look reasonable. For instance, two frequent words denoting events: the noun *toename* (increase) and the verb *stijgen* (increase) appear to have sensible patients, shown in Table 5.7.

5.8 Conclusion

We described a knowledge-rich approach to the interoperable extraction of event data from text, expressed in different ways and across different languages. We use a shared representation formats for seven different languages and shared modules for the semantic processing of the text. Ontological implications from a single shared ontology are inserted in the text using wordnets and WSD. We used a pattern-matching module to extract event-participant relations from text running over these ontological statements. We evaluated the system on sentences of a single document, which showed promising results.

In the near future, we will extend the evaluation of the system to other types of text and more languages. We will also exploit many more options to use semantic constraints on interpreting sequences that have not been exploited yet. We will especially investigate more precise ways in which WSD can be combined with the task to come to an interpretation of the textual elements that makes sense. Finally, we will work on the more complex ways in which the different events fit together. So far we consider each event as separate but the examples in Sect. 5.2 showed that event descriptions overlap to a high degree.

Acknowledgements The KYOTO project is co-funded by EU – FP7 ICT Work Programme 2007 under Challenge 4 – Digital libraries and Content, Objective ICT-2007.4.2 (ICT-2007.4.4): Intelligent Content and Semantics (challenge 4.2). The Asian partners from Tapei and Kyoto are funded from national funds. This work has been also supported by Spanish project KNOW-2 (TIN2009-14715-C04-01).

References

1. Agirre, E., Soroa, A.: Personalizing PageRank for word sense disambiguation. In: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL '09), pp. 33–41. Association for Computational Linguistics, Stroudsburg (2009)
2. Baker, C.F., Fillmore, C.J., Lowe, J.B.: The berkeley framenet project. In: Proceedings of the 17th International Conference on Computational Linguistics, COLING '98, vol. 1, pp. 86–90. Association for Computational Linguistics, Stroudsburg (1998)
3. Bontcheva, K., Wilks, Y.: Automatic report generation from ontologies: the MIAKT approach. In: Nineth International Conference on Applications of Natural Language to Information Systems (NLDB'2004). Manchester (2004)
4. Bosma, W.E., Vossen, P., Soroa, A., Rigau, G., Tesconi, M., Marchetti, A., Monachini, M., Aliprandi, C.: Kaf: a generic semantic annotation format. In: Proceedings of the GL2009 Workshop on Semantic Annotation, Pisa (2009)
5. Brooke, J.: SUS: a quick and dirty usability scale. In: Jordan, P.W., Thomas, B., Weerdmeester, B.A., McClelland, A.L. (eds.) Usability Evaluation in Industry. CRC (1996)
6. Fellbaum, C.: WordNet: an Electronical Lexical Database. The MIT Press, Cambridge (1998)
7. Gangemi, A., Guarino, N., Masolo, C., Oltramari, A.: Interfacing WordNet with DOLCE: towards OntoWordNet. In: Ontology and the Lexicon, pp. 36–52. Cambridge University Press, Cambridge/New York (2010)
8. Guarino, N., Welty, C.: Evaluating ontological decisions with ontoclean. Commun. ACM **45**(2), 61–65 (2002)
9. Hicks, A., Herold, A.: Evaluating ontologies with rudify. In: Dietz J.L.G. (ed.) Proceedings of the 2nd International Conference on Knowledge Engineering and Ontology Development (KEOD'09), pp. 5–12. INSTICC Press (2009)
10. Huyhn, D., Karger, D., Miller, R.: Exhibit: lightweight structured data publishing. ACM 978-1-59593-654-7/07/0005. MIT Computer Science and Artificial Intelligence Laboratory (2007)
11. Ide, N., Romary, L.: Outline of the international standard linguistic annotation framework. In: Proceedings of the ACL 2003 Workshop on Linguistic Annotation: Getting the Model Right (LingAnnot 03), vol. 19, pp. 1–5. Association for Computational Linguistics, Stroudsburg (2003)

12. Izquierdo, R., Suarez, A., Rigau, G.: Exploring the automatic selection of basic level concepts. In: Angelova, G., Bontcheva, K., Mitkov, R., Nicolov, N., Nikolov, N. (eds.) International Conference Recent Advances in Natural Language Processing, Borovets, pp. 298–302 (2007)
13. Kaiser, K., Miksch, S.: Information extraction a survey. Tech. rep., Vienna University of Technology. Institute of Software Technology and Interactive Systems (2005)
14. Majid, A., Boster, J.S., Bowerman, M.: The cross-linguistic categorization of everyday events: a study of cutting and breaking. Cognition **109**, 235–250 (2008)
15. Masolo, C., Borgo, S., Gangemi, A., Guarino, N., Oltramari, A.: Wonderweb deliverable d18: Ontology library. Tech. rep., ITSC-CNR, Trento (2003)
16. Mizoguchi, R., Sunagawa, E., Kozaki, K., Kitamura, Y.: The model of roles within an ontology development tool: hozo. Appl. Ontol. **2**, 159–179 (2007)
17. Niles, I., Pease, A.: Linking lexicons and ontologies: mapping wordnet to the suggested upper merged ontology. In: Proceedings of the 2003 International Conference on Information and Knowledge Engineering (IKE 03), Las Vegas, pp. 23–26. CSREA Press, Las Vegas (2003)
18. Peshkin, L., Pfeffer, A.: Bayesian information extraction network. In: Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI'03), pp. 421–426. Morgan Kaufmann, San Francisco (2003)
19. Pustejovsky, J., Lee, K., Bunt, H., Romary, L.: Iso-timeml: an international standard for semantic annotation. In: Chair, N.C.C., Choukri, K., Maegaard, B., Mariani, J., Odijk, J. Piperidis, S., Rosner, M., Tapias, D. (eds.) Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10). European Language Resources Association (ELRA), Valletta (2010)
20. Putnam, H.: The meaning of 'meaning'. Minn. Stud. Philos. Sci. **7**, 131–193 (1975)
21. Vossen, P., Rigau, G.: Division of semantic labor in the global wordnet grid. In: Proceedings of Global WordNet Conference (GWC'2010), Mumbay (2010)
22. Vossen, P., Agirre, E., Calzolari, N., Fellbaum, C., Shu-Kai Hsieh, Chu-Ren Huang, Isahara, H., Kanzaki, K., Marchetti, A., Monachini, M., Neri, F., Raffaelli, R., Rigau, G., Tesconi, M., VanGent, J.: KYOTO: a system for mining, structuring, and distributing knowledge across languages and cultures. In: Proceedings of the 4th Global WordNet Conference (GWC'08), University of Szeged. Szeged, Hungary (2008)