KYOTO: an open platform for mining facts

Piek Vossen
VU University Amsterdam
p.vossen@let.vu.nl

German Rigau
Eneko Agirre
Aitor Soroa
University of the Basque Country
german.rigau/e.agirre/a.soroa@ehu.es

Monica Monachini
Roberto Bartolini
Istituto di Linguistica Computazionale, CNR
monica.monachini/roberto.bartolini@ilc.cnr.it

Abstract

This document describes an open text-mining system that was developed for the Asian-European project KYOTO. The KYOTO system uses an open text representation format and a central ontology to enable extraction of knowledge and facts from large volumes of text in many different languages. We implemented a semantic tagging approach that performs off-line reasoning. Mining of facts and knowledge is achieved through a flexible pattern matching module that can work in much the same way for different languages, can handle efficiently large volumes of documents and is not restricted to a specific domain. We applied the system to an English database on estuaries.

1 Introduction

Traditionally, Information Extraction (IE) is the task of filling template information from previously unseen text which belongs to a predefined domain (Peshkin & Pfeffer 2003). Most systems in the Message Understanding Conferences (MUC, 1987-1998) and the Automatic Content Extraction program (ACE)[1] use a pipeline of tools to achieve this, ranging from sophisticated NLP tools (like deep parsing) to shallower text-processing (e.g. FASTUS (Appelt 1995)).

Standard IE systems are based on language-specific pattern matching (Kaiser & Miksch 2005), where each pattern consists of a regular expression and an associated mapping from syntactic to logical form. In general, the approaches can be categorized into two groups: (1) the Knowledge Engineering approach (Appelt et al.1995), and (2) the learning approach, such as AutoSlog (Appelt et al. 1993), SRV (Freitag 1998), or RAPIER (Califf & R. Mooney 1999). Another important system is GATE (Cunningham et al.2002), which is a platform for creating IE systems. It uses regular expressions, but it can also use ontologies to perform semantic inferences to constrain linguistic patterns semantically. The use of ontologies in IE is an emerging field (Bontcheva & Wilks 2004): linking text instances with elements belonging to the ontology, instead of consulting flat gazetteers.

The major disadvantage of traditional IE systems is that they focus on satisfying precise, narrow, pre-specified requests from small homogeneous corpora (e.g., extract information about terrorist events). Likewise, they are not flexible, are limited to specific types of knowledge and need to be built by knowledge engineers for each specific application and language. In fact most text mining systems are developed for a single domain and a single language, and are not able to handle knowledge expressed in different languages or expressed and conceptualized differently across cultures.

In this paper we describe an open platform for text-mining or IE that can be applied to many different languages in the same way using an open text representation system and a central on-

[1] http://www.itl.nist.gov/iad/mig//tests/ace

Proceedings of the 6th Workshop on Ontologies and Lexical Resources (Ontolex 2010), pages 1–10,
Beijing, August 2010

tology that is shared across languages. Ontological implications are inserted in the text through off-line reasoning and ontological tagging. The events and facts are extracted from large amounts of text using a flexible pattern-matching module, as specified by profiles which comprise ontological and shallow linguistic patterns. The system is developed in the Asian-European project KYOTO[2].

In the next section, we describe the general architecture of the KYOTO system. In section 3, we specify the knowledge structure that is used. Section 4, describes the off-line reasoning and ontological tagging. In section 5, we describe the module for mining knowledge from the text that is enriched with ontological statements. Finally in section 6, we describe the first results of applying the system to databases on Estuaries.

2 KYOTO overview

The KYOTO project allows communities to model terms and concepts in their domain and to use this knowledge to apply text mining on documents. The knowledge cycle in the KYOTO system starts with a set of source documents produced by the community, such as PDFs and websites. Linguistic processors apply tokenization, segmentation, morpho-syntactic analysis and semantic processing to the text in different languages. The semantic processing involves the detection of named-entities (persons, organizations, places, time-expressions) and determining the meaning of words in the text according to the given wordnet.

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, Bosma et al 2009). This format incorporates standardized proposals for the linguistic annotation of text and represents them in an easy-to-use layered structure, which is compatible with the Linguistic Annotation Framework (LAF, Ide and Romary 2003). 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 basic output, when needed (Bosma et al. 2009, Vossen et al. 2010). All modules in KYOTO draw their input from these structures. In fact, the word-sense disambiguation process is carried out to the same KAF annotation in different languages and is therefore the same for all the languages (Agirre et al. 2009). In the current system, there are processors for English, Dutch, Italian, Spanish, Basque, Chinese and Japanese.

The KYOTO system proceeds in 2 cycles (see Figure 1). In the 1st cycle, the **Tybot** (Term Yielding Robot) extracts the most relevant terms from the documents. The Tybot is another generic program that can do this for all the different languages in much the same way. The terms are organized as a structured hierarchy and, wherever possible, related to generic semantic databases, i.e. wordnets for each language. In the left part of Figure 1, we show those terms in the input document and their classification in wordnet. Terms in italics are present in the original wordnet, while underlined terms correspond to terms which were not in the original wordnet but were automatically discovered and linked to wordnet by Tybots. Straight terms correspond to hyperonyms in wordnet that do not necessarily occur in the text but are linked to ontological classes. The result of this 1st cycle is a domain wordnet for the target language.

The 2nd cycle of the system involves the actual extraction of factual knowledge from the documents by the **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. Since the semantics is defined through the ontology, it is possible to detect similar data across documents in different languages, even if expressed differently. In Figure 1, we give an example of a conceptual pattern that relates organisms that live in habitats. The Kybot can combine morpho-syntactic and semantic patterns. When a match is detected, the instantiation of the pattern is saved in a formal representation, either in KAF or in RDF. 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.

[2] Http://www.kyoto-project.eu

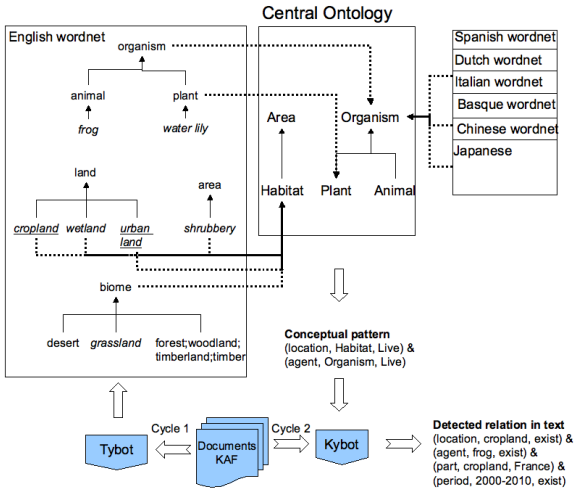

Figure 1: Two Cycles of processing in KYOTO

3 Ontological and lexical background knowledge

As a semantic background model, we defined a 3-layered knowledge architecture following the principle of the division of labour (Putnam 1975). In this model, the ontology does not need to be the central hub for all terms in a domain in all languages. Following the division of labour principle, we can state that a computer does not need to distinguish between instances of a European Tree Frog and a Glass Tree frog. We assume that rigid concepts (as defined by Guarino and Welty 2002) are known to the domain experts and do not need to be defined formally in the ontology but can remain in the available background resources, such as databases with millions of species. Terms in the documents are mostly non-rigid, e.g. *endangered frogs*, *invasive frogs*. Such non-rigid terms refer to instances of species in contextual circumstances. The processes and states are the important pieces of information that matter to the users and are useful for mining text. The model therefore distinguishes between background vocabularies, domain terms, wordnets and the central ontology. The background vocabularies are automatically aligned to wordnet, where we assume that hyponymy relations to rigid synsets in wordnet declare those subconcepts as rigid subtypes too, without the necessity to include them in the ontology. For non-rigid terms, we defined a set of mapping relations to the ontology through which we express their non-rigid involvement in these processes and states. Likewise, the ontology has been extended with processes and states for the domain and verbs and adjectives have been mapped to be able to detect expressions in text.

The 3-layered knowledge model combines the efforts from 3 different communities:

1. Domain experts in social communities that continuously build background vocabularies;
2. Wordnet specialists that define the basic semantic model for general concepts for a language
3. Semantic Web specialists that define top-level and domain-specific ontologies that capture formal definitions of concepts;

We formalized the relations between these repositories so that they can developed separately but combined within KYOTO to form a coherent and formal model.

3.1 Ontology

The KYOTO ontology currently consists of 1149 classes divided over three layers. The top layer is based on DOLCE (DOLCE-Lite-Plus version 3.9.7, Masolo et al 2003) and OntoWordNet. This layer of the ontology has been modified for our purposes (Herold et. al. 2009). The second layer consists of so-called Base Concepts (BCs) derived from various wordnets (Vossen 1998, Izquierdo et al. 2007). Examples of BCs are: *building, vehicle, animal, plant, change, move, size, weight*. The BCs are those synsets in WordNet 3.0 that have the most relations with other synsets in the wordnet hierarchies and are selected in a way that ensures complete coverage of the nominal and verbal part of WordNet. This has been completed for the nouns (about 500 synsets). The ontology has also been adapted to include important concepts in the domain. Special attention has been paid to represents the processes (**perdurants**) in which objects (**endurants**) of the domain are involved and qualities they may have. This is typically the information that is found in documents on the environment. We thus added 40 new event classes for representing important verbs (e.g. *pollute*, *absorb*, *damage*, *drain*) and 115 new qualities and quality-regions for representing important adjectives (e.g. *airborne*, *acid*, (*un*)*healthy*, *clear*). The full

ontology can be downloaded from the KYOTO website, free for use. A considerable set of general verbs and adjectives (relevant for for the domain) have then been mapped to ontological classes: 189 verbal synsets and 222 adjectival synsets.

The 500 nominal BCs are connected to the complete WordNet hierarchy, whereas the 189 verbs represent 5,978 more specific verbal synsets and the 222 adjectives represent 1,081 adjectival synsets through the wordnet relations.

This basic ontology and the mapping to WordNet are used to model the shared and language-neutral concepts and relations in the domain. Instances are excluded from the ontology. Instances will be detected in the documents and will be mapped to the ontology through instance to ontology relations (see below). Likewise, we make a clear separation between the ontological model and the instantiation of the model as described in the text.

3.2 Wordnet to ontology mappings

In addition to the ontology, we have wordnets for each language in the domain. In addition to the regular synset to synset relations in the wordnet, we will have a specific set of relations for mapping the synsets to the ontology, which are all prefixed with *sc_* standing for synset-to-concept. We differentiate between rigid and non-rigid concepts in the wordnets through the mapping relations:

- sc_**equivalenceOf**: the synset is fully equivalent to the ontology Type & inherits all properties; the synset is Rigid
- sc_ **subclassOf**: the synset is a proper subclass of the ontology Type & inherits all properties; the synset is Rigid
- sc_**domainOf**: the synset is not a proper subclass of the ontology Type & is not disjoint (therefore orthogonal) with other synsets that are mapped to the same Type either through sc_subclassOf or sc_domainOf; the synset is non-Rigid but still inherits all properties of the target ontology Type; 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 (Mizoguchi et al. 2007).[3]
- sc_**participantOf**: instances of the concept (denoted by the synset) participate in some endurant, where the specific role relation is indicated by the playRole mapping.
- sc_**hasState**: instances of the concept are in a particular state which is not essential and can be changed. There is no need to represent the role for a stative perdurant.

This model extends existing WordNet to ontology mappings. For instance, in the SUMO to Wordnet mapping (Niles and Pease 2003), only the *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. In our model, we separate the linguistically and culturally specific vocabularies from the shared ontology while using the ontology to interface the concepts used by the various communities.

Using these mapping relations, we can express that the synset for *duck* (which has a hypernym relation to the synset *bird*, which, in its turn, has an equivalence relation to the ontology class *bird*) is thus a proper subclassOf the ontology class *bird*:

wn:duck hypernym *wn:bird*
wn:bird sc_equivalenceOf *ont:bird*

For a concept such as *migratory bird*, which is also a hyponym of *bird* in wordnet but not a proper subclass as a non-rigid concept, we thus create the following mapping:

wn:migratory bird
→ sc_domainOf *ont:bird*
→ sc_playRole *ont:done-by*
→ sc_participantOf *ont:migration*

This mapping indicates that the synset is used to refer to instances of endurants (not subclasses!), where the domain is restricted to birds. Furthermore, these instances participate in the process of

[3] Some terms involve more than one role, e.g. gas-powered-vehicle. Secondary participants are related through sc_**hasCoParticipant** and sc_**playCoRole** mappings.

migration in the role of *done-by*. The properties of the process migration are further defined in the ontology, which indicates that it is a active-change-of-location done-by some endurant, going from a source, via a path to some destination. The mapping relations from the 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.

For implied non-essential states, we use the *sc_hasState* relation to express that a synset such as *wild dog* refers to instances of dogs that life in the *wild* but can stop being *wild*:

wn:wild dog → sc_domainOf *ont:dog*
wn:wild dog → sc_hasState *ont:wild*

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 (e.g. *to pollute, to reduce, wild*). However, domain specific text contain many new nominal terms that refer to domain-specific processes and states, e.g. *air pollution, nitrogen pollution, nitrogen reduction*. These terms are equally relevant as their counter-parts that refer to endurants involved in similar processes, e.g. *polluted air, polluting nitrogen or reduced nitrogen*. We therefore use the reverse participant and role mappings to be able to define such terms for processes as subclasses of more general processes involving specific participants in a specified role:

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*

Further mapping relations are described in the documentation on the KYOTO website. Through the mapping relations, we can keep the ontology relatively small and compact whereas we can still define the richness of the vocabularies of languages in a precise way. The classes in the ontology can be defined using rich axioms that model precise implications for inferencing. The wordnet to synset mappings can be used to define rather basic relations relative to the given ontology that still captures the semantics of the terms. The term definitions capture both relevance and perspective (those relations that matter from the point of the view of the term), on the one hand, and some semantics with respect to the concepts that are involved and their (role) relation on the other hand. Likewise, the KYOTO system can model the linguistic and cultural diversity of languages in a domain but at the same time keep a firm anchoring to a basic and compact ontology.

3.3 Domain wordnet

We selected 3 representative documents on estuaries to extract relevant terms for the domain using the Tybot module. The terms have been related through structural relations, e.g. *nitrogen pollution* is a hyponym of *pollution*, and through WordNet synsets that are assigned through WSD of the text. We extracted 3950 candidate terms form the KAF representations of the documents. Most of these are nouns (2818 terms). The nominal terms matched for 40% with wordnet synsets, the verbs and adjectives for 98% and 85% respectively. For the domain wordnet, we restricted ourselves to the nouns. From the new nominal terms, environmentalists selected 390 terms that they deem to be important. These terms are connected to parent terms, which ultimately are connected to wordnet synsets. The final domain wordnet contains 659 synsets: 197 synsets from the generic wordnet and 462 new synsets connected to the former. The domain wordnet synsets got 990 mappings to the ontology, using the relations described in the previous section. There are 86 synsets that have a sc_domainOf mapping, indicating that they are non-rigid. Note that hyponyms of these synsets are also non-rigid by definition. These non-rigid synsets have complex mappings to processes and states in which they are involved. The domain wordnet can be downloaded from the KYOTO website, free for use.

```
<term lemma="pollution" pos="N" tid="t13444" type="open">
  <externalReferences>
   <externalRef reference="eng-30-00191142-n" reftype="baseConcept" resource="wn30g"/>
   <externalRef reference="Kyoto#change-eng-3.0-00191142-n" reftype="sc_subClassOf" resource="ontology">
      <externalRef reftype="SubClassOf" reference="DOLCE-Lite.owl#contamination_pollution"/>
      <externalRef reftype="SubClassOf" reference="DOLCE-Lite.owl#accomplishment" status="implied"/>
      <externalRef reftype="SubClassOf" reference="DOLCE-Lite.owl#event" status="implied"/>
      <externalRef reftype="SubClassOf" reference="DOLCE-Lite.owl#perdurant" status="implied"/>
      <externalRef reftype="DOLCE-Lite.owl#part" reference="DOLCE-Lite.owl#perdurant" status="implied"/>
      <externalRef reftype="DOLCE-Lite.owl#specific-constant-constituent" reference="DOLCE-Lite.owl#perdurant"
status="implied"/>
      <externalRef reftype="DOLCE-Lite.owl#has-quality" reference="DOLCE-Lite.owl#temporal-quality" status="implied"/>
      <externalRef reftype="SubClassOf" reference="DOLCE-Lite.owl#spatio-temporal-particular" status="implied"/>
      <externalRef reftype="DOLCE-Lite.owl#participant" reference="DOLCE-Lite.owl#endurant" status="implied"/>
      <externalRef reftype="DOLCE-Lite.owl#has-quality" reference="DOLCE-Lite.owl#temporal-location_q" status="im-
plied"/>
    <externalRef reftype="SubClassOf" reference="DOLCE-Lite.owl#particular" status="implied"/>
    </externalRef>
  </externalReferences>
</term>
```

Figure 2: An example of an OntoTagged output

```
<kprofile>
 <variables>
  <var name="x" type="term" pos="N"/>
  <var name="y" type="term"
       lemma="produce | generate | release | ! create"/>
  <var name="z" type="term"
       reference="DOLCE-Lite.owl#contamination_pollution"
       reftype="SubClassOf"/>
 </variables>
 <relations>
  <root span="y"/>
  <rel span="x" pivot="y" direction="preceding"/>
  <rel span="z" pivot="y" direction="following"/>
 </relations>
 <events>
  <event target="$y/@tid" lemma="$y/@lemma" pos="$y/@pos"/>
  <role target="$x/@tid" rtype="agent" lemma="$x/@lemma"/>
  <role target="$z/@tid" rtype="patient"lemma="$z/@lemma"/>$
 </events>
</kprofile>
```

Figure 3: An example of a Kybot profile

```
<kybotOut>
 <doc name="11767.mw.wsd.ne.onto.kaf">
  <event eid="e1" lemma="generate" pos="V" target="t3504"/>
  <role rid="r1" lemma="industry" rtype="agent" target="t3493" pos="N" event="e1"/>
  <role rid="r2" lemma="pollution" rtype="patient" target="t3495" pos="N" event="e1"/>
 </doc>
 <doc name="16266.mw.wsd.ne.onto.kaf">
  <event eid="e2" lemma="release" pos="V" target="t97"/>
  <role rid="r3" lemma="fuel" rtype="agent" target="t96" pos="N" event="e2"/>
  <role rid="r4" lemma="exhaust_gas" rtype="patient" target="t101" pos="V" event="e2"/>
 </doc>
</kybotOut>
```

Figure 4: An example of a Kybot output

4 Off-line reasoning and ontological tagging

The ontological tagging represents the last phase in the KYOTO Linguistic Processor annotation pipeline. It consists of a three-step module devised to enrich the KAF documents with knowledge derived from the ontology. For each synset connected to a term, the first step adds the Base Concepts to which the synset is related through the wordnet taxonomical relations. Then, through the synset to ontology mapping, it adds the corresponding ontology type with appropriate relations. Once each synset is specified as to its ontology type, the last ontotagging step inserts the full set of ontological implications that follow from the explicit ontology. The explicit ontology is a new data structure consisting of a table with all ontology nodes and all ontological implications expressed. The main purpose is to optimize

the performance of the mining module over large quantities of documents. The advantage for Kybots from ontotagging are many. First of all, they are able to run and apply pattern-matching to Base Concepts and ontological classes rather than just to words or synsets. Moreover, by making explicit the implicit ontological statements, Kybots are able to find the same relations hidden in different expressions with different surface realizations: *fish migration, migratory fish, migration of fish, fishes that migrate,* that directly or indirectly express the same relations. With ontotagging, they share the same ontological implications which will allow Kybots to apply the same patterns and perform the extraction of facts. The implications will be represented in the same way across different languages, thus facilitating cross-lingual extraction of facts. Lastly, ontotagging is a kind of off-line ontological reasoning: without doing reasoning over concepts, Kybots substantially improve their performance. Figure 2 shows the result of onto-tagging for the term *pollution*.

5 Event and fact extraction

Kybots (Knowledge Yielding Robots) are computer programs that use the mined concepts and the generic concepts already connected to the language wordnets and the KYOTO ontology to extract actual concept instances and relations in KAF documents. Kybots incorporate technology for the extraction of relationships, either eventual or not, relative to the general or domain concepts already captured by the Tybots. That is, the extraction of factual knowledge is being carried out by the Kybot server by processing Kybot profiles on the linguistically enriched documents.

Kybots are defined following a declarative format, the so called *Kybot profiles*, which describe general morpho-syntactic and semantic conditions on sequences of terms. Profiles are compiled to generate the Kybots, which scan over KAF documents searching for the patterns and extract the relevant information from each matching.

Linguistic patterns include morphologic constraints and also semantic conditions the matched terms must hold. 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 Section 3.2. Thus, it is possible to detect similar eventual information across documents in different languages, even if expressed differently.

5.1 Example of a Kybot Profile

Kybot Profiles are described using XML syntax. Figure 3 presents an example of a profile. Kybot profiles consist of three main parts:

•*Variable declaration* (`<variables>` element): In this section the search entities are defined. The example defines three variables: **`x`** (denoting terms whose part-of-speech is noun), **`y`** (which are terms whose lemma is “release”, “produce” or “generate” but not “create”) and **`z`** (terms linked to the ontological endurant “DOLCE-Lite.owl#contamination_pollution”, meaning ``being contaminated with harmful substances").

•*Declarations of the relations among variables* (`<rel>` element): specify the relations among the previously defined variables. The example profile specifies **`y`** as the main pivot, and states that variable **`x`** must be preceding variable **`y`** in the same sentence, and that variable **`z`** must be following variable **`y`**. Thus, the Kybot will search for patterns like '**`x`** → **`y`** → **`z`**' in a sentence.

•*Output template* (`<events>` element): describes the output to be produced on every matching. In the example, each match generates a new event targeting term **`y`**, which becomes the main term of the event. It also fills two roles of the event, the 'agent' role filled by term **`x`** and 'patient' role, filled by **`z`**.

Figure 4 presents the output of the Kybot when applied against the benchmark documents. The Kybot output follows the stand-off architecture when producing new information, and it thus forms a new KAF layer on the original documents.

6 Experimental results

We applied the KYOTO system and resources to English documents on estuaries. We collected 50 URLs for two English estuaries: the Humber Estuary in Hull (UK) and the Chesapeake Bay estuary in the US and for background documents on bird migration, sedimentation, habitat destruction, and climate change. In addition to the webpages, we extracted 815 PDF files from the sites. In total, 4625 files have been extracted. All

the documents have been processed by the linguistic processor for English, which generated KAF representations for all the documents. From this database, 3 documents were selected for benchmarking.

The documents were processed by applying multiword tagging, word-sense-disambiguation, named-entity-recognition and the ontological tagging to the 3 documents and to the complete database; This was done twice: once without the domain model and once with the domain model. We thus created 4 datasets: 3 benchmark documents processed with and without the domain model; the complete database processed with and without the domain model.

Furthermore, we created Kybot profiles based on the type of information represented in the domain model. We applied the Kybots to all 4 data sets. We generate the following data files through an WN-LMF export of the domain wordnet:

1. a set of domain multiwords for the multiword tagger
2. an extension of the lexicon and the graph of concepts that is used by the WSD module
3. an extension of the wordnet-to-ontology mappings for the ontotagger

In addition, we constructed mapping lists for all WordNet 3.0 synsets to Base Concepts and to adjective and verbs that are matched to the ontology. These mappings provide the generic conceptual model based on wordnet and on the ontology.

Table 1 shows the effects of using the domain model for the first 3 modules. We can see that the domain model has a clear effect on the multiword detection in the 3 evaluation documents. Using the domain model, 600 multiwords have been detected, against 145 with just the generic wordnet. This is obvious since the terms are extracted from the same documents. However, when applying it to the complete database, we see that still over 2,300 more multiwords have been detected using the domain wordnet. Note that the domain wordnet has only 97 multiwords and the generic wordnet has 19,126 multiwords. So 0.5% of the multiwords in the domain wordnet add 1.5 times more multiword tokens in the database. The third row specifies the number of synsets that have been assigned. We can see that for the domain model almost 400 more synsets have been detected. In the case of the full estuary database, we see that relatively few more have been detected, almost 1,500 while the database is 80 times as big. If we look more closely at the numbers of actual domain synsets detected, we see the following results. In the benchmark documents 637 (or 5%) of the synsets is a domain wordnet synset, whereas 5,353 synsets are domain synsets in the full estuary database, which is only 0.52%. Note that in KAF multiwords are represented both as a single terms and in terms of their elements. The WSD module assigns synsets to both. The domain model can thus only add synsets compared to the processing without the domain.

Finally, if we look at the named-entity-recognition module, we see a slight negative effect for the detection of named-entities due to the domain model. The named-entity-recognition module does not consider the elements of multiwords but just the multiword terms as a whole. Grouping terms as multiwords thus leads to less named-entities being detected. This is not necessarily a bad things, since the detection heavily over-generates and could have now more precision.

	bench mark documents (3)		estuary documents (4742)	
	No Domain	Domain	No Domain	Domain
terms	22,204	22,204	2,419,839	2,419,839
multiwords	145	600	4,389	6,671
synsets	12,526	12,910	1,021,598	1,023,017
ne location	158	126	41,681	40,714
ne date	67	66	10,288	10,233

Table 1: Statistics on processing the estuary documents with and without domain model

	bench mark documents (3)				estuary documents (4272	
	No Domain		Domain		Domain	
ontology references	555,677		576,432		48,708,300	
implied ontology referer	457,332	82.30%	474,916	82.39%	40,523,452	83.20%
direct ontology referenc	53,178	9.57%	54,769	9.50%	4,377,814	8.99%
domain synset to ontolo	45,167	8.13%	46,747	8.11%	3,807,034	7.82%

Table 2: Ontological implications for the four data sets

Table 2 shows the effect of inserting ontological implications into the text representation.For the benchmark documents, we see that more than half a million ontological implications have been inserted. Of these, 82% are implied references, that are extracted from the explicit ontology on the basis of a direct mapping to the ontology. About 8% of the mappings are synset-to-ontology mappings (sc) and 9.5% are mappings representing the subclass hierarchy. The differences between using the domain model and not-using the domain model are minimal. For the complete database, the implications are 80 times as much but the proportions are similar.

Table 3 shows the type of sc-relations that occur. Obviously, sc_subClassOf and sc_equivalentOf are the most frequent. Nevertheless, we still find about 500 mappings that present the participation in a process or state.

```
   30  reftype="sc_playCoRole"
   32  reftype="sc_hasCoParticipant"
   42  reftype="sc_partOf"
   59  reftype="sc_stateOf"
   92  reftype="sc_playRole"
   94  reftype="sc_hasRole"
   97  reftype="sc_participantOf"
  105  reftype="sc_hasParticipant"
  128  reftype="sc_domainOf"
  169  reftype="sc_hasState"
  312  reftype="sc_hasPart"
 3637  reftype="sc_equivalentOf"
42048  reftype="sc_subClassOf"
```

Table 3: Type of relations for the wordnet to ontology mappings using the domain model

The table clearly shows the impact of role relations that are encoded in the domain wordnet. When we extract the mappings for the files without the domain model (ony using the mappings to the generic wordnet), we get only equivalence and subclass mappings.

Finally to complete the knowledge cycle, we created a few Kybot profiles for extracting events from the onto-tagged documents. As an initial test, 3 profiles have been created:

1. events of destruction
2. destructions of locations
3. destruction of objects

Using these profiles, we extracted 211 events from the 3 benchmark documents with 396 roles. The profiles are created to run over the ontological types inserted by the ontotagger, e.g. restricted to events and change_of_integrity. Despite the generality of the profiles, we still see a clear signature of the domain in the output. This is a good indication that we will be able to extract valuable events from the data, even though the ontotagger generates a massive amount of implications. Especially events that combine multiple roles appear to give rich information. For example, the following sentence:

"One of the greatest challenges to restoration is continued population growth and development, which destroys forests, wetlands and other natural areas"

yielded the following output:

```
  <event target="t1471" lemma="destroy" pos="V"
eid="e74"/>
  <role target="t1477" rtype="patient" lemma="area"
pos="N" event="e74" rid="r138"/>
  <role target="t1472" rtype="patient"
lemma="forest" pos="N" event="e74" rid="r151"/>
  <role target="t1469" rtype="actor" lemma="devel-
opment" pos="N" event="e74" rid="r180"/>
```

Running the full set of profiles on the complete database with almost 60 million ontological statements took about 2 hours. This shows that our approach is scalable and efficient.

7 Conclusions

In this paper, we described an open platform for text-mining using wordnets and a central ontology. The system can be used across different languages and can be tailored to mine any type of conceptual relations. It can handle semantic implications that are expressed in very different linguistic expressions and yield systematic output. As future work, we will carry out benchmarking and testing of the mining of events, both for English and for the other languages in the KYOTO project.

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 Contsent 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

Agirre, E., & Soroa, A. (2009) Personalizing PageRank for Word Sense Disambiguation. Proceedings of the 12th EACL, 2009. Athens, Greece.

Agirre, E., Lopez de Lacalle, O., & Soroa, A. (2009) Knowledge-based WSD and specific domains: performing over supervised WSD. Proceedings of IJCAI. Pasadena, USA. http://ixa.si.ehu.es/ukb

Álvez J., Atserias J., Carrera J., Climent S., Laparra E., Oliver A. and Rigau G. (2008) Complete and Consistent Annotation of WordNet using the Top Concept Ontology. Proceedings of LREC'08, Marrakesh, Morroco. 2008.

Appelt Douglas E., Jerry R. Hobbs, John Bear, David Israel, Megumi Kameyama, Andrew Kehler, David Martin, Karen Myers and Mabry Tyson. Description of the FASTUS System Used for MUC-6. In Proceedings of MUC-6, pages 237–248. San Mateo, Morgan Kaufmann, 1995.

Auer A., C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak and Z. Ives. DBpedia: A Nucleus for a Web of Open Data. In Proceedings of the International Semantic Web Conference (ISWC), volume 4825 of Lecture Notes in Computer Science, pages 722-735. 2007.

Bosma, W., Vossen, P., Soroa, A. , Rigau, G., Tesconi, M., Marchetti, A., Monachini, M., & Apiprandi, C. (2009) KAF: a generic semantic annotation format. In Proceedings of the 5th International Conference on Generative Approaches to the Lexicon Sept 17-19, 2009, Pisa, Italy.

Fellbaum, C. (Ed.) (1998) WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.

Freitag, D. (1998) Information extraction from html: Application of a general machine learning approach. In Proceedings of the Fifteenth National Conference on Artificial Intelligence, 1998.

Gangemi A., Guarino N., Masolo C., Oltramari A., Schneider L. (2002) Sweetening Ontologies with DOLCE. Proceedings of EKAW. 2002

Ide, N. and L. Romary. 2003. Outline of the inter- national standard Linguistic Annotation Framework. In *Proceedings of ACL 2003 Workshop on Linguistic Annotation: Getting the Model Right*, pages 1–5.

Izquierdo R., Suárez A. & Rigau G. Exploring the Automatic Selection of Basic Level Concepts. Proceedings of RANLP'07, Borovetz, Bulgaria. September, 2007.

Masolo, C., Borgo, S., Gangemi, A., Guarino, N. & Oltramari, A. (2003) WonderWeb Deliverable D18: Ontology Library, ISTC-CNR, Trento, Italy.

Mizoguchi R., Sunagawa E., Kozaki K. & Kitamura Y. (2007 A Model of Roles within an Ontology Development Tool: Hozo. Journal of Applied Ontology, Vol.2, No.2, 159-179.

Niles, I. & Pease, A. (2001) Formal Ontology in Information Systems. Proceedings of the international Conference on Formal Ontology in Information Systems – Vol. 2001 Ogunquit, Maine, USA

Niles, I. and A. Pease. Linking lexicons and ontologies: Mapping WordNet to the Suggested Upper Merged Ontology. In Proc. IEEE IKE, pages 412–416, 2003.

Vossen, P. (Ed.) (1998) EuroWordNet: a multilingual database with lexical semantic networks for European Languages. Kluwer, Dordrecht.

Vossen P., W. Bosma, E. Agirre, G. Rigau, A. Soroa (2010) A full Knowledge Cycle for Semantic Interoperability. Proceedings of the 5th Joint ISO-ACL/SIGSEM Workshop on Interoperable Semantic Annotation, (ICGL 2010) Hong Kong, 2010.