



KYOTO (ICT-211423) Intelligent Content and Semantics
Knowledge **Y**ielding **O**ntologies for **T**ransition-Based **O**rganization
<http://www.kyoto-project.eu/>

Event and Fact Mining

German Rigau
IXA group, UPV/EHU

Final Review
April 8th, 2011, Berlin, Germany



ICT-211423

Knowledge Mining in Kyoto

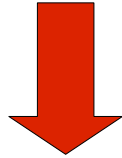
- Concept mining (Tybot)
 - Extract terms and relations in a language
 - Map the terms to an existing wordnet
 - Ontologize terms to concepts and axioms
- Fact mining (**Kybot**)
 - Define **morpho-syntactic** and **semantic patterns** in text
 - Extract events from text
 - Collect events and extract facts
- For all languages!
- **KAF** (Kyoto Annotation Format) is the input of both:
 - Tybot: term extraction
 - Kybot: fact extraction

Outline

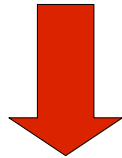
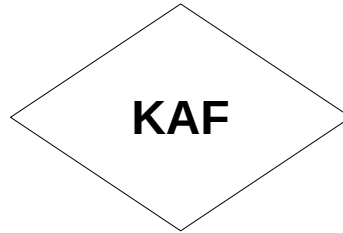
- Kyoto CORE for fact extraction
- Knowledge Architecture
- Mining module
- Kybot evaluation & benchmarking
- Future development

Fact Mining: Kybots

Tropical terrestrial species populations declined
by 55 per cent from 1970 to 2003



- + Linguistic Processing: POS, chunks, dependencies, ...
- + Semantic Processing: WSD ($\Rightarrow$ WN $\Rightarrow$ ontology)



- + **Kybot profiles**: morphosyntactic + semantic patterns
- + Mining Module: Events / Facts

Tropical terrestrial species **populations declined**
by **55** per cent **from 1970 to 2003**

KAF

- Based on current ISO proposals
- Language-neutral annotation of text,
 - concepts, facts,...
 - Multilingual
- Interoperable across linguistic processors
 - KAF is the basis for integration
- Flexible and extendible

Linguistic Processors

- KAF (Kyoto Annotation Format)
 - English: **Synthema**
 - Dutch: **VUA**
 - Italian: **Synthema**
 - Basque: **EHU**
 - Spanish: **EHU**
 - Chinese: **AS**
 - Japanese: **NICT**
- MW detection: **VUA**
- Word Sense Disambiguation module (UKB): **EHU**
- NE Tagger: **Irion**
- OntoTagger: **CNR-ILC, EHU**

Linguistic Processors

- KAF XML files include sections for:
 - Word forms
 - Terms / Items
 - Chunks: grouping of sequences of terms
 - Dependencies: syntactic relations between terms
 - WSD: WN senses of the term
 - Ontological references of the term:
 - Base Concepts
 - Explicit ontology
 - Events
 - Locations, Time expressions
 - ...

Fact Mining: Kybot profiles

- Kybot profiles consist of:
 - Morpho-syntactic conditions
 - LPs outcomes
 - Semantic conditions:
 - WordNets + Ontologies
 - **Inferencing** on WN & ontology !
 - Output Template
 - Event / Fact descriptions

Fact Mining: Kybot profiles

- For each sentence :
 - **IF** Morpho-syntactic Conditions match **and**
 - Semantic Conditions hold
 - **THEN**
 - generate the Output Template
- How to make **efficient** inferencing on WN & ontology?
 - ... while processing very large volumes of KAF
 - WN => Nominal and Verbal Base Concepts !
 - Ontology => Explicit Ontology !
 - **Off-line inferencing** !

Knowledge Architecture

- Modeling **domain** knowledge ...
 - for **seven languages**
 - each one encoding diverse **phenomena**
 - ... *migratory bird ... birds that migrate ...*
 - ... *migratory path / pattern ...*
 - ... *migration of ducks ...*
 - general and specialized **terminology**
 - ... *footprint ... greenhouse gas ...*
 - ... *Humber estuary ...*
 - ... *SAC features – littoral and sub-tidal ...*
 - ... *SPA ...*
 - ... *cape teal ... anas capensis ...*
 - ... *Yellow-billed Pintail ...*
 - ...

File Edit View History Bookmarks Tools Help

http://en.wikipedia.org/wiki/Anas

Most Visited Release Notes Error Fedora Project Red Hat Free Content

W Anas - Wikipedia, the free ... linux which java installed - ... Pellet 2.0.1 Release Downl... Gmail - Inbox (4524) - griga...

Try Beta Log in / create account

Anas

From Wikipedia, the free encyclopedia

For other uses, see [Anas \(disambiguation\)](#).

Anas is a [genus](#) of [dabbling ducks](#). It includes [mallards](#), [wigeons](#), [teals](#), [pintails](#) and [shovelers](#) in a number of [subgenera](#). Some authorities prefer to elevate the subgenera to genus rank^[1]. Indeed, as the [moa-nalos](#) are very close to this clade and may have evolved later than some of these lineages, it is rather the absence of a thorough review than lack of necessity that this genus is rather [over-lumped](#).

Contents [\[hide\]](#)

- Systematics
 - Species
 - Fossil record
- Footnotes
- References
- External links

Systematics [\[edit\]](#)

The [phylogeny](#) of this genus is one of the most confounded ones of all living birds. Research is hampered by the fact the radiation of the two major groups of *Anas* - the [teals](#) and mallard groups -; took place in a very short time and fairly recently, roughly in the mid-late [Pleistocene](#). Furthermore, [hybridization](#) may have long played a [major role](#) in *Anas* evolution, with within-subgenus hybrids regularly and between-subgenus hybrids not infrequently being fully fertile^[1] see also [Mariana Mallard](#). The relationships between species are much obscured by this fact, and [mtDNA sequence](#) data is of dubious value in resolving their relationships^[2]; on the other hand, [nuclear DNA](#) sequences evolve too slowly to resolve the [phylogeny](#) of the subgenus *Anas* for example.

Some major [clades](#) can be discerned. For example, that the traditional [subgenus](#) *Anas*, the mallard group, forms a

Anas

Female [Mallard](#) (*Anas platyrhynchos*) with brood of young, a typical member of this genus.

Scientific classification

Kingdom:	Animalia
Phylum:	Chordata
Class:	Aves
Subclass:	Neornithes
Infraclass:	Neognathae
Superorder:	Galloanserae
Order:	Anseriformes

navigation

- Main page
- Contents
- Featured content
- Current events
- Random article

search

Go Search

interaction

- About Wikipedia
- Community portal
- Recent changes
- Contact Wikipedia
- Donate to Wikipedia
- Help

toolbox

- What links here
- Related changes
- Upload file
- Special pages

Done

Knowledge Repositories for the domain

- **Term database**: 100,000 terms per language
- **DBPedia**: 2.6 million things
- **GeoNames**: 8 million geographical names
- **Species 2000**: 2.1 million species
- **Wordnets** for 7 languages:
 - about 50,000 to 120,000 synsets per language
 - Domain WN: ~2,000 concepts
- **Ontologies**: SUMO, DOLCE-Lite, SIMPLE
 - Kyoto ontology 3.1: 1500 classes
- ...

Knowledge Integration in KYOTO

- Should all knowledge be stored in the **central ontology**?
 - The knowledge is (still) **too large**
 - The knowledge to be stored is **too diverse**
 - Different types of knowledge require **different inferencing capabilities**

Knowledge Integration in KYOTO

- A model of **division of labour** (along the lines of Putnam 1975) in which knowledge is stored in **3 layers**:
 - Vocabularies, term databases, etc. (SKOS)
 - WordNet (WN-LMF)
 - Ontology (OWL-DL)
- **Mapping relations** that support the division of labour
 - language-specific conceptualizations
- Each layer supports different types of **inferencing**
 - SparQL queries
 - Graph algorithms (UKB, SSID+)
 - Formal reasoning (OWL-DL reasoners, FACT++)

KYOTO Knowledge Model

ONTOLOGY

~ thousands of **types** : **MOVE**

Extension of DOLCE-Lite including *Base Concepts*

synset2TypeRelations

WORDNET

~ hundreds of thousands of **concepts**: **<migratory#a>**

EquivalenceRelation

VOCABULARY

~millions of **terms**: **migratory#a**

Language-dependant

Automatic selection of Base Concepts

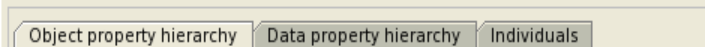
- **Base Concepts** are the result of a compromise between two **conflicting** principles of characterization:
 - Represent as many concepts as possible
 - Represent as many features as possible
- Base Concepts typically occur in the middle of semantic hierarchies

Automatic selection of Base Concepts

freq.	#rel	synset
2338	18	00017954-n group 1,grouping 1
0	19	05962976-n social group 1
729	37	05997592-n organisation 2,organization 1
30	10	06002286-n establishment 2,institution 1
15	12	06023733-n faith 3,religion 2
62	5	06024357-n Christianity 2, church 1 ,Christian church 1
11	14	00001740-n entity 1,something 1
51	29	00009457-n object 1,physical ob ject 1
1	39	00011937-n artifact 1,artefact 1
68	63	03431817-n construction 3,structure 1
50	79	02347413-n building 1,edifice 1
0	11	03135441-n place of worship 1,house of prayer 1
59	19	02438778-n church 2 ,church building 1
25	20	00017487-n act 2,human action 1,human activity 1
611	69	00261466-n activity 1
2	5	00662816-n ceremony 3
0	11	00663517-n religious ceremony 1,religious ritual 1
243	7	00666638-n service 3,religious service 1,divine service 1
11	1	00666912-n church 3 ,church service 1

WordNet to Ontology mappings

- By using the Base Concepts as an abstraction layer, all WN synsets have been connected to the Ontology
 - 297 **nominal** Base Concepts
 - 578 **verbal** Base Concepts
- WN hierarchy for nouns and verbs
- Non hierarchical relations for **adjectives**
 - **Morpho-semantic links** from WN



The screenshot shows a window titled 'has-destination' with a list of properties. The 'entails' property is selected, and its details are shown on the right. The details include a 'has-destination' property with a value of 'entails' and a 'has-path' property with a value of 'entails'.

Annotations: bird genus-eng-3.0-01507175-n

comment "(genus of birds)"

label

Equivalent classes 

Superclasses +

genus-eng-3.0-08108972-n

Inferred anonymous superclasses

```
has-quality some (binary_quality
  or indefinite_quality
  or measurable_quality)
```

- particular
- and enduring

- or perdurant
- or quality

physical-object Inherited from spatio-temporal-particular

and proper-part only (member-of exactly 1 Thing)

part only **endurant**

- specific-constant-constituent **only** endurant

● participant-in some perdurant

- specific-constant-constituent **only** physical-endurant

- part **only** physical-endurant
- has quality **only** physical quality

has-quality only physical-quality

has-quality some physical-quality

- has quality **some** physical-quality

Wordnet ontology relations

Rigid vs. Non-rigid

Rigid

- Synset:Endurant; Synset:Perdurant; Synset:Quality:
- sc_equivalenceOf

Non-rigid:

- Synset:Role; Synset:Endurant
- sc_domainOf: range of ontology types that restricts a role
- sc_playRole: role that is being played

Rigidity can be detected automatically (**Rudify**, 80% precision, IAG 80%) and is stored in wordnets as attributes to synsets

Wordnet ontology-relations

sc_equivalenceOf

sc_subclassOf

sc_domainOf

sc_playRole

sc_participantOf

sc_hasState

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

Lexicalization of process-related concepts

{obstruct, obturate, impede, occlude, jam, block, close up}Verb, English

-> sc_equivalenceOf **ObstructionPerdurant**

{obstruction, obstructor, obstructer, impediment, impedimenta}Noun, English

-> sc_domainOf **PhysicalObject**

-> sc_playRole **ObstructingRole**

{migration birds}Noun, English

-> sc_domainOf **Bird**

-> sc_playRole **MigratorRole**

{migration}Verb, English

-> sc_equivalenceOf **MigrationProcess**

{migration area}Noun, English

-> sc_domainOf **PhysicalObject**

-> sc_playRole **TargetRole**

Lexicalization of process-related concepts

{create, produce, make}Verb, English

-> sc_equivalenceOf **ConstructionProcess**

{artifact, artefact}Noun, English

-> sc_domainOf **PhysicalObject**

-> sc_playRole **ConstructedRole**

{kunststof}Noun, Dutch // lit. *artifact substance*

-> sc_domainOf **AmountOfMatter**

-> sc_playRole **ConstructedRole**

{meat}Noun, English

-> sc_domainOf **Cow, Sheep, Pig**

-> sc_playRole **EatenRole**

{名肉, 食物, 餐}Noun, Chinese

-> sc_domainOf **Cow, Sheep, Pig, Rat, Mole, Monkey**

-> sc_playRole **EatenRole**

{طعام, لحم, غذاء}Noun, Arabic

-> sc_domainOf **Cow, Sheep**

-> sc_playRole **EatenRole**

WordNet to Ontology mappings

{07312616} (n) **migration** (the periodic passage of groups of animals (especially birds or fishes) from one region to another for feeding or breeding)

eng-30-07312616-n sc_subClassOf

Kyoto#happening__occurrence__occurrent__natural_event-eng-3.0-07283608-n

eng-30-07312616-n sc_subClassOf

Kyoto#move-eng-3.0-01855606-v

{01857093} (v) **migrate** (move periodically or seasonally) "birds migrate in the Winter";
"The workers migrate to where the crops need harvesting"

eng-30-01857093-v sc_subClassOf

Kyoto#happening__occurrence__occurrent__natural_event-eng-3.0-07283608-n

eng-30-01857093-v sc_subClassOf

Kyoto#move-eng-3.0-01855606-v

{02129007} (adj) **migratory** (used of animals that move seasonally) "migratory birds"

eng-30-02129007-a sc_subClassOf

Kyoto#happening__occurrence__occurrent__natural_event-eng-3.0-07283608-n

eng-30-02129007-a sc_subClassOf

Kyoto#move-eng-3.0-01855606-v



WordNet to Ontology mappings

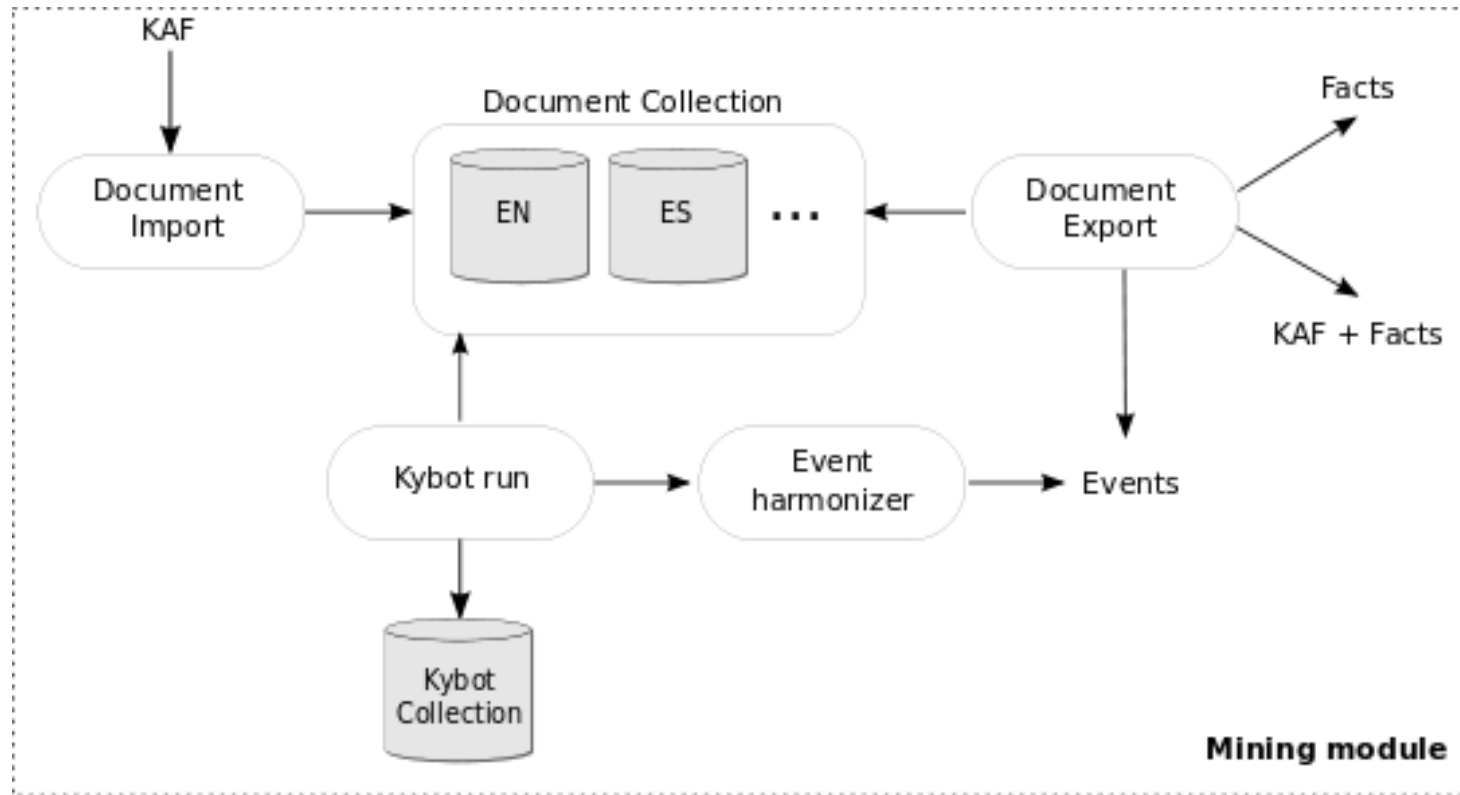
Kyoto#move-eng-3.0-01855606-v SubClassOf Kyoto#move-eng-3.0-01855606-v inherited
Kyoto#move-eng-3.0-01855606-v SubClassOf Kyoto#change_of_location__movement_11-eng-3.0-00280586-n
Kyoto#move-eng-3.0-01855606-v SubClassOf Kyoto#verb_motion
Kyoto#move-eng-3.0-01855606-v SubClassOf DOLCE-Lite.owl#perdurant inherited
Kyoto#move-eng-3.0-01855606-v merged.owl#pertinent-quality DOLCE-Lite.owl#spatial-location_q inherited
Kyoto#move-eng-3.0-01855606-v SubClassOf Kyoto#change-eng-3.0-00191142-n inherited
Kyoto#move-eng-3.0-01855606-v merged.owl#initial-quality DOLCE-Lite.owl#space-region inherited
Kyoto#move-eng-3.0-01855606-v merged.owl#end-quality DOLCE-Lite.owl#space-region inherited
Kyoto#move-eng-3.0-01855606-v **DOLCE-Lite.owl#participant** DOLCE-Lite.owl#endurant inherited
Kyoto#move-eng-3.0-01855606-v **Kyoto#has-path** DOLCE-Lite.owl#particular inherited
Kyoto#move-eng-3.0-01855606-v **Kyoto#has-source** DOLCE-Lite.owl#particular inherited
Kyoto#move-eng-3.0-01855606-v **Kyoto#has-destination** DOLCE-Lite.owl#particular inherited
Kyoto#move-eng-3.0-01855606-v DOLCE-Lite.owl#has-quality DOLCE-Lite.owl#temporal-location_q inherited
Kyoto#move-eng-3.0-01855606-v SubClassOf DOLCE-Lite.owl#spatio-temporal-particular inherited
Kyoto#move-eng-3.0-01855606-v DOLCE-Lite.owl#has-quality DOLCE-Lite.owl#temporal-quality inherited
Kyoto#move-eng-3.0-01855606-v DOLCE-Lite.owl#part DOLCE-Lite.owl#perdurant inherited
Kyoto#move-eng-3.0-01855606-v SubClassOf DOLCE-Lite.owl#accomplishment inherited
Kyoto#move-eng-3.0-01855606-v DOLCE-Lite.owl#specific-constant-constituent DOLCE-Lite.owl#perdurant inherited
Kyoto#move-eng-3.0-01855606-v SubClassOf DOLCE-Lite.owl#particular inherited
Kyoto#move-eng-3.0-01855606-v SubClassOf DOLCE-Lite.owl#event inherited



KYOTO pipelines & WSD service

- English / Italian
 - http://wiki.ilc.cnr.it/kyoto_demo/index.php
- English / Spanish / Basque
 - <http://ixa2.si.ehu.es/demokaf/demokaf.pl>
- WSD by evocation
 - <http://xmlgroup.iit.cnr.it/demos/WSD/>

Mining Module Architecture



- KAF (ontotagged) Documents stored in XML DB
- Kybots are stored in XML documents (files)
- Kybots are executed using XQueries on the XML DB

Kybot application

- User uploads documents to the collection
- User applies a series of Kybots to documents
 - Or a subset of docs (ex. only a language)
- Kybots create new events and facts
- Also, keep track of which kybot created which fact

Kybot profiles

- Self descriptive (for manual Kybot creation)
- Pattern-matching like, plus many capabilities.
- Use XML syntax to define the kybots
- Efficient
 - Able to manage thousands of KAF documents

Kybot profiles

- Powerful expressions
- POS
- Lemma
- Senses, Base Concepts
- Ontological references
- Suffix/prefix expressions
- Conjunction, disjunction, optionality
- Negation
- Chunks
- Not in between
- Predicate-filler Kybots

Kybot profiles

```
<?xml version="1.0" encoding="utf-8"?>
```

```
<Kybot id="Generate_Pollution">
```

```
<variables>
```

```
  <var name="X" type="term" pos="N"/>
```

```
  <var name="Y" type="term" lemma="release | produce | generate | ! create"/>
```

```
  <var name="Y" pos="V"/>
```

```
  <var name="Z" type="term" lemma="*pollution | pollutant | contaminant"/>
```

```
</variables>
```

```
<relations>
```

```
  <root span="X"/>
```

```
  <rel span="Y" pivot="X" direction="following"/>
```

```
  <rel span="Z" pivot="Y" direction="following"/>
```

```
</relations>
```

```
<events>
```

```
  <event target="$Y/@tid" lemma="$Y/@lemma" pos="$Y/@pos"/>
```

```
  <role target="$X/@tid" rtype="source" lemma="$X/@lemma" pos="$X/@pos"/>
```

```
  <role target="$Z/@tid" rtype="patient" lemma="$Z/@lemma" pos="$Z/@pos"/>
```

```
</events>
```

```
</Kybot>
```



Kybot profiles

```
<?xml version="1.0" encoding="utf-8"?>
```

```
<Kybot id="Generate_Pollution">
```

```
<variables>
```

Variables

```
  <var name="X" type="term" pos="N"/>
```

```
  <var name="Y" type="term" lemma="release | produce | generate | ! create"/>
```

```
  <var name="Y" pos="V"/>
```

```
  <var name="Z" type="term" lemma="*pollution | pollutant | contaminant"/>
```

```
</variables>
```

```
<relations>
```

```
  <root span="X"/>
```

```
  <rel span="Y" pivot="X" direction="following"/>
```

```
  <rel span="Z" pivot="Y" direction="following"/>
```

```
</relations>
```

```
<events>
```

```
  <event target="$Y/@tid" lemma="$Y/@lemma" pos="$Y/@pos"/>
```

```
  <role target="$X/@tid" rtype="source" lemma="$X/@lemma" pos="$X/@pos"/>
```

```
  <role target="$Z/@tid" rtype="patient" lemma="$Z/@lemma" pos="$Z/@pos"/>
```

```
</events>
```

```
</Kybot>
```



Kybot profiles

```
<?xml version="1.0" encoding="utf-8"?>
```

```
<Kybot id="Generate_Pollution">
```

```
<variables>
```

```
  <var name="X" type="term" pos="N"/>
```

```
  <var name="Y" type="term" lemma="release | produce | generate | ! create"/>
```

```
  <var name="Y" pos="V"/>
```

```
  <var name="Z" type="term" lemma="*pollution | pollutant | contaminant"/>
```

```
</variables>
```

```
<relations>
```

```
  <root span="X"/>
```

```
  <rel span="Y" pivot="X" direction="following"/>
```

```
  <rel span="Z" pivot="Y" direction="following"/>
```

```
</relations>
```

Relations

```
<events>
```

```
  <event target="$Y/@tid" lemma="$Y/@lemma" pos="$Y/@pos"/>
```

```
  <role target="$X/@tid" rtype="source" lemma="$X/@lemma" pos="$X/@pos"/>
```

```
  <role target="$Z/@tid" rtype="patient" lemma="$Z/@lemma" pos="$Z/@pos"/>
```

```
</events>
```

```
</Kybot>
```



Kybot profiles

```
<?xml version="1.0" encoding="utf-8"?>
```

```
<Kybot id="Generate_Pollution">
```

```
<variables>
```

```
  <var name="X" type="term" pos="N"/>
```

```
  <var name="Y" type="term" lemma="release | produce | generate | ! create"/>
```

```
  <var name="Y" pos="V"/>
```

```
  <var name="Z" type="term" lemma="*pollution | pollutant | contaminant"/>
```

```
</variables>
```

```
<relations>
```

```
  <root span="X"/>
```

```
  <rel span="Y" pivot="X" direction="following"/>
```

```
  <rel span="Z" pivot="Y" direction="following"/>
```

```
</relations>
```

```
<events>
```

```
  <event target="$Y/@tid" lemma="$Y/@lemma" pos="$Y/@pos"/>
```

```
  <role target="$X/@tid" rtype="source" lemma="$X/@lemma" pos="$X/@pos"/>
```

```
  <role target="$Z/@tid" rtype="patient" lemma="$Z/@lemma" pos="$Z/@pos"/>
```

```
</events>
```

Output Template

```
</Kybot>
```



Kybot profiles: Output

```
<kybotOut>
  <doc shortname="1534.mw.wsd.ne.onto.kaf">
    <event target="t886" lemma="generate" pos="V" eid="e1"/>
    <role target="t884" rtype="source" lemma="watershed" .../>
    <role target="t892" rtype="patient" lemma="pollution" .../>
  </doc>
  <doc shortname="17795.mw.wsd.ne.onto.kaf">
    <event target="t9690" lemma="release" pos="V" eid="e1"/>
    <role target="t9691" rtype="patient" lemma="pollutant" .../>
    <role target="t9678" rtype="source" lemma="fuel" .../>
    <role target="t9680" rtype="source" lemma="heating" .../>
    <role target="t9681" rtype="source" lemma="machinery" .../>
    <role target="t9683" rtype="source" lemma="equipment" .../>
    <role target="t9686" rtype="source" lemma="household" .../>
    <role target="t9688" rtype="source" lemma="business" .../>
  </doc>
</kybotOut>
```

Complex profiles

```
<Kybot id="generic_kybot-accomplishment-affectORimpact-physical-endurant">
  <variables>
    <var name="v1" type="term" lemma="! can" reftype="SubClassOf"
      reference="DOLCE-Lite.owl#accomplishment"/>
    <var name="v1" type="term" lemma="! do"/>
    <var name="vnot1" type="term" pos="V | P | D"/>
    <var name="v2" type="term" pos="V" lemma="affect | impact"/>
    <var name="v3" type="term" pos="N" reftype="SubClassOf"
      reference="DOLCE-Lite.owl#physical-endurant"/>
    <var name="vnot2" type="term" pos="V | P"/>
  </variables>

  <relations>
    <root span="v3"/>
    <rel span="v1" pivot="v2" direction="preceding" notInBetween="vnot1"/>
    <rel span="v2" pivot="v3" direction="preceding" notInBetween="vnot2"/>
  </relations>

  <events>
    <event target="$v2/@tid"/>
    <role target="$v1/@tid" rtype="simple-cause-of"/>
    <role target="$v3/@tid" rtype="patient"/> </events> </Kybot>
```



Kybot profiles: Output (simplified)

```
<kybotOut>
<doc shortname="11767.mw.wsd.ne.onto.kaf">
  <event eid="e1" target="t779" lemma="impact" pos="V" />
  <role rid="r1" event="e1" target="t778" lemma="pollution" pos="N"
    rtype="simple-cause-of" />
  <role rid="r2" event="e1" target="t782mw" lemma="chesapeake bay" pos="N"
    rtype="patient" />
  <role rid="r3" event="e1" target="t785" lemma="tributary" pos="N"
    rtype="patient" />
  <event eid="e2" target="t1644" lemma="affect" pos="V" />
  <role rid="r4" event="e2" target="t1643" lemma="snowfall" pos="N"
    rtype="simple-cause-of" />
  <role rid="r5" event="e2" target="t1646mw" lemma="water flow" pos="N"
    rtype="patient" />
  <event eid="e3" target="t5045" lemma="affect" pos="V" />
  <role rid="r6" event="e3" target="t5042" lemma="water" pos="N"
    rtype="simple-cause-of" />
  <role rid="r7" event="e3" target="t5048" lemma="level" pos="N"
    rtype="patient" />
</doc>
</kybotOut>
```

Predicate-filler Kybots: Kybots and Ontology

- Profiles combine syntactic patterns and ontological information
- For example:

X (noun) << Y (verb)
participant-in(event:Y, filler:X)

<externalRef reftype="Kyoto#active-participant-in"
reference="Kyoto#protection-eng-3.0-00817680-n"/>

One of the major concerns of the **Linconsire's Wildlife
Crime Officer** is the **protection** of the estuary habitats.

<externalRef reference="Kyoto#protection-eng-3.0-00817680-n"
reftype="SubclassOf"/>

Predicate-filler Kybots

```
<Kybot id="generic_kybot">
  <variables>
    <var name="X" type="term" pos="N"/>
    <var name="Y" type="term" pos="V"/>
  </variables>

  <relations>
    <root span="Y"/>
    <rel span="X" pivot="Y" direction="preceding"/>
    <predicate pred="DOLCE-Lite.owl#participant-in"
               event="Y" filler="X"/>
  </relations>

  <events>
    <event target="$Y/@tid" lemma="$Y/@lemma" pos="$Y/@pos"/>
    <role target="$X/@tid" lemma="$X/@lemma" pos="$X/@pos"
          rtype="participant"/>
  </events>
</Kybot>
```

Kybot output

```
<kybotOut>
  <doc name="11614.mw.wsd.ne.onto.kaf">
    <event eid="e1" target="t1718" lemma="protect" pos="N"/>
    <role rid="r1" event="e1" target="t1715"
      rtype="participant" lemma="crime_officer" pos="N"/>
    ...
  </doc>
</kybotOut>
```

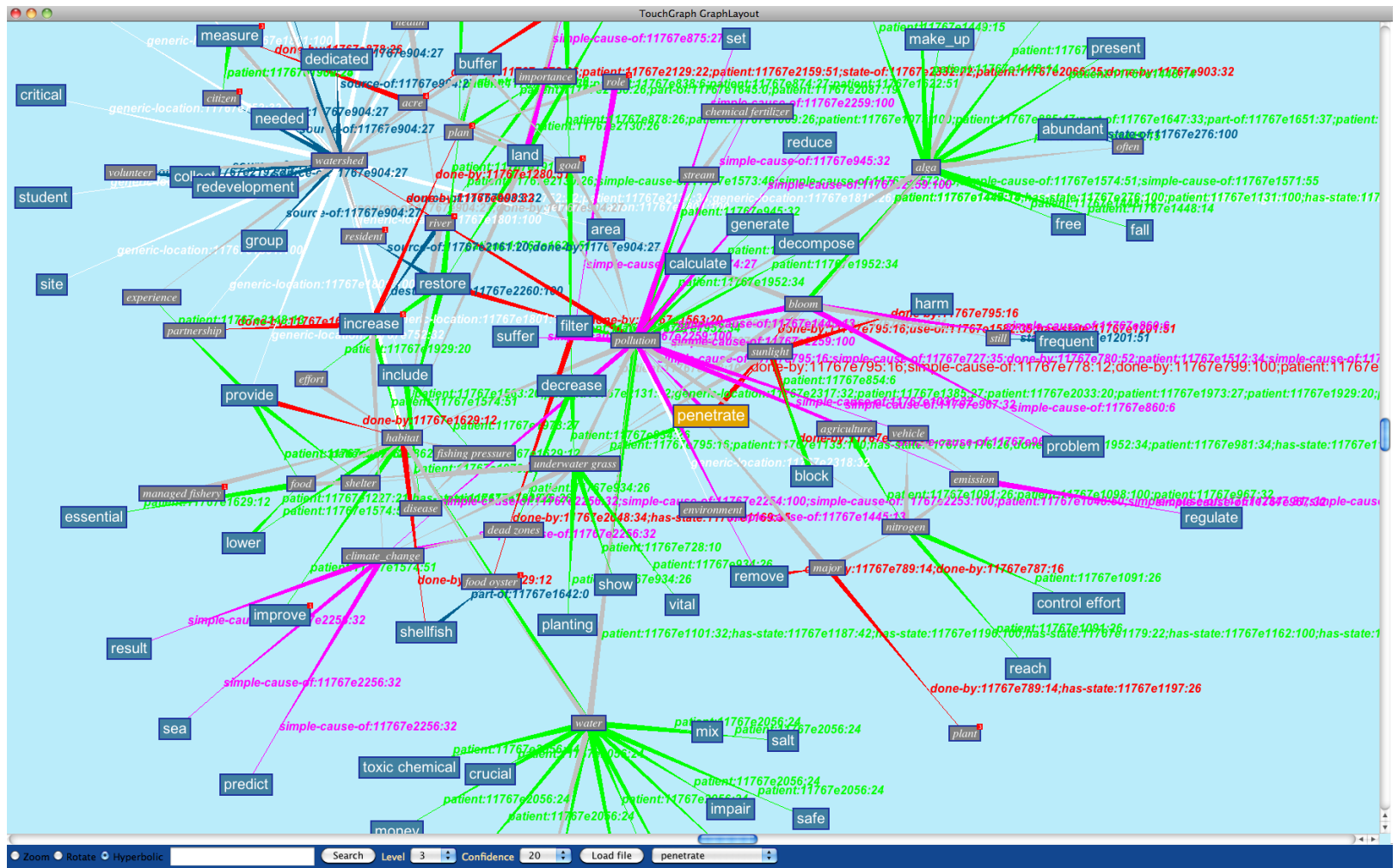

Event Harmonizer

- Group events and facts
 - Refer to same term and synset
- Locate events/roles in space/time
 - NER module: identify locations and dates in documents
 - Apply heuristics to events/roles to associate best location/date

Kybot output with dates/locations

```
<doc shortname="23452.mw.wsd.ne.onto.kaf">
  <event eid="e1" target="t723" lemma="graze" pos="V"
    synset="eng-30-00669762-v" rank="0.0329727">
    <place countryCode="GB" countryName="United Kingdom" latitude="52.2"
      longitude="-2.6666667" name="Humber" timezone="Europe/London">
      <span id="t721"/>
    </place>
    <dateInfo dateISO="1999" lemma="1999">
      <span id="t527"/>
    </dateInfo>
  </event>
  <role rid="r1" event="e1" target="t731" lemma="outer estuary" pos="N"
    rtype="generic-location" synset="eng-30-09225146-n" rank="0.19" >
    <place ...>...</place>
    <dateInfo ...>...</dateInfo>
  </role>
  ...
</doc>
```

Collecting events/roles



Kybot Evaluation & Benchmarking

- Profiles
- Benchmarking
- Project internal evaluation
 - Gold-standard
 - Error analysis
 - Effect of WSD
 - Effect of best profiles
 - Effect of domain modelling
- Open competition
- Transferring Kybots to another language
- Transferring Kybots to another domain

Profiles

- Currently we have **261** generic profiles
 - Manually developed
 - Search for generic ontological relations
 - “accomplishment affects/impacts accomplishment”
 - “accomplishment of biological-object”
 - ...
- Specific profiles for extracting implicit events in compounds
 - “migratory bird” evokes a migration event
 - “crab exploitation” has 'crabs' as patients
 - etc.

Performance

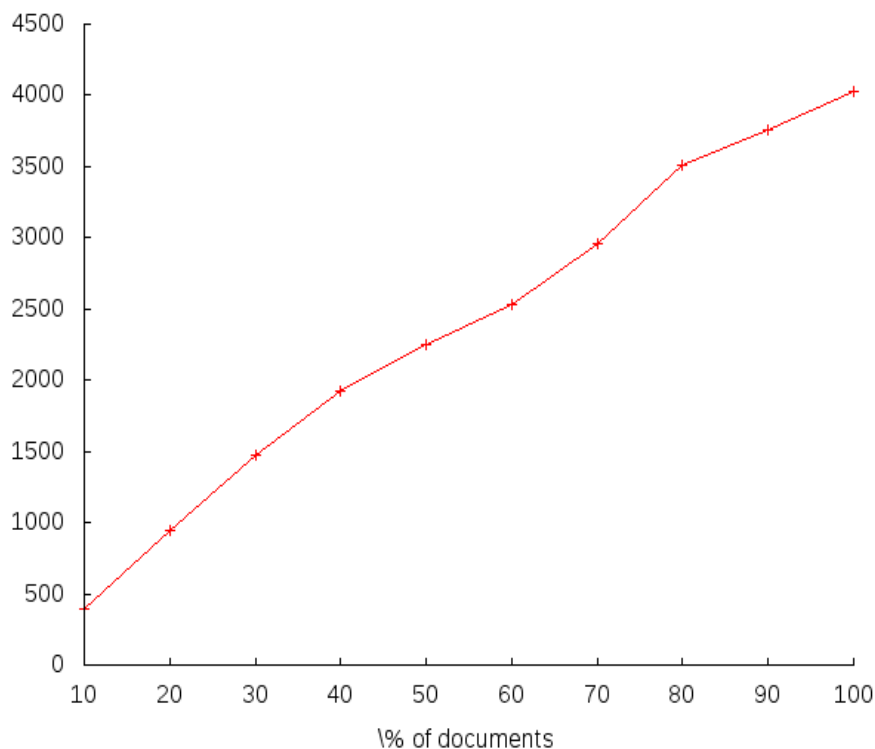
- Running times on medium size and big corpora
 - Subset of 60 profiles
 - Two corpus
 - **Benchmark** corpus
 - 21,721 words
 - 706,646 external references
 - **Estuary** corpus
 - ~3 million terms
 - ~60 million external references

	Benchmark	Estuary
N. events	2,936	185,012
Time	119s	16,112s

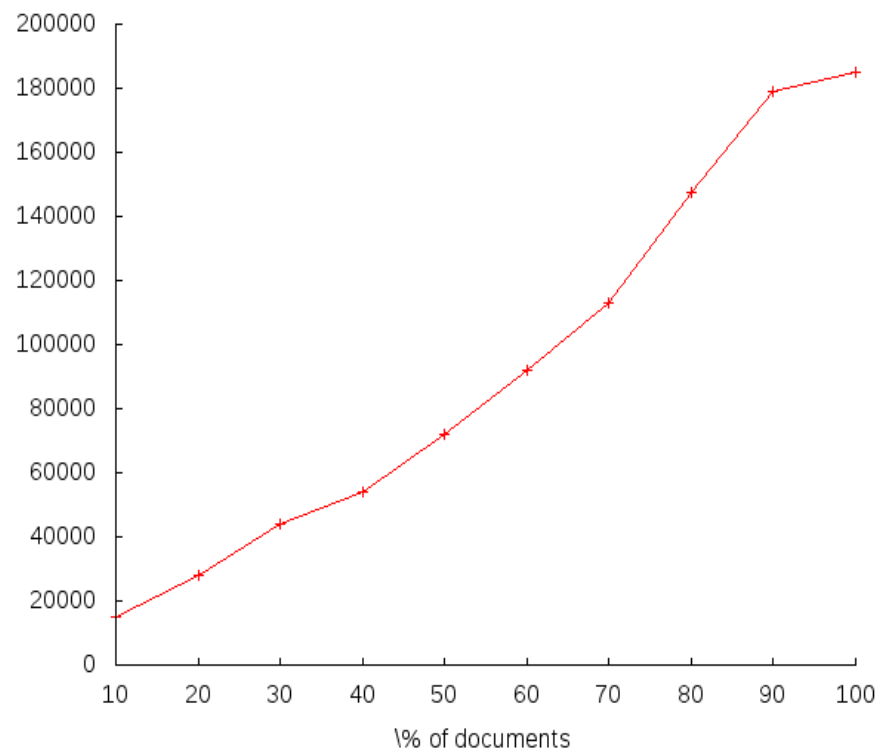
Performance varying corpus size

- Measure performance with different size corpora
- On average, 20 facts per second

Time



N. of facts



Kybot output evaluation

- Create **gold standard**
 - Choose one document:
[www.acb-online.org/pubs/BayBarometer2008 Web.pdf](http://www.acb-online.org/pubs/BayBarometer2008%20Web.pdf)
 - Manually annotate events/roles
 - Convert events/roles to triplets
 - **KafAnnotator**
 - 127 sentences, 1,416 tokens
 - Annotate 353 triplets (201 unique events)
- Run Kybot profiles and measure precision/recall.

Kybot output evaluation

	Ignored Relation	Patient
Nr. correct	306	115
Precision	0.09	0.03
Recall	0.86	0.33

Table 3: Baseline of chunk heads in the same sentence.

	Ignored relation	All relations
Nr. correct	222	174
Precision	0.49	0.32
Recall	0.63	0.49

Table 4: Generic processing with 261 profiles

Kybot output evaluation

- **Error Analysis**
 - More / better information from the **parser**
 - POS errors
 - Complex grammatical extructures
 - Compositionality
 - Some annotation errors / discrepancies
 - Some knowledge errors
 - Missing WordNet concepts
 - Wrong ontology class or mapping
 - Missing profiles

Kybot output evaluation

- Effects of **WSD**
 - Excluding low scoring concepts when there is a **choice** (multiple interpretations)
 - event / role
 - different relations

Kybot output evaluation

WSD threshold	#triplets	# in scope	# correct	P.	R.	F1
0	1816	548	174	0.32	0.49	0.39
10	1551	500	169	0.34	0.48	0.40
20	1469	479	167	0.35	0.47	0.40
30	1399	470	167	0.36	0.47	0.41
40	1351	461	166	0.36	0.47	0.41
50	1272	446	164	0.37	0.46	0.41
60	1226	434	164	0.38	0.46	0.42
70	1214	429	162	0.38	0.46	0.41
80	1206	427	161	0.38	0.46	0.41
90	1190	426	161	0.38	0.46	0.41
100	1085	377	148	0.39	0.42	0.41
manual	605	364	141	0.39	0.40	0.39

Table 7: Generic processing with different WSD thresholds.

- Better recall than manual WSD !

Kybot output evaluation

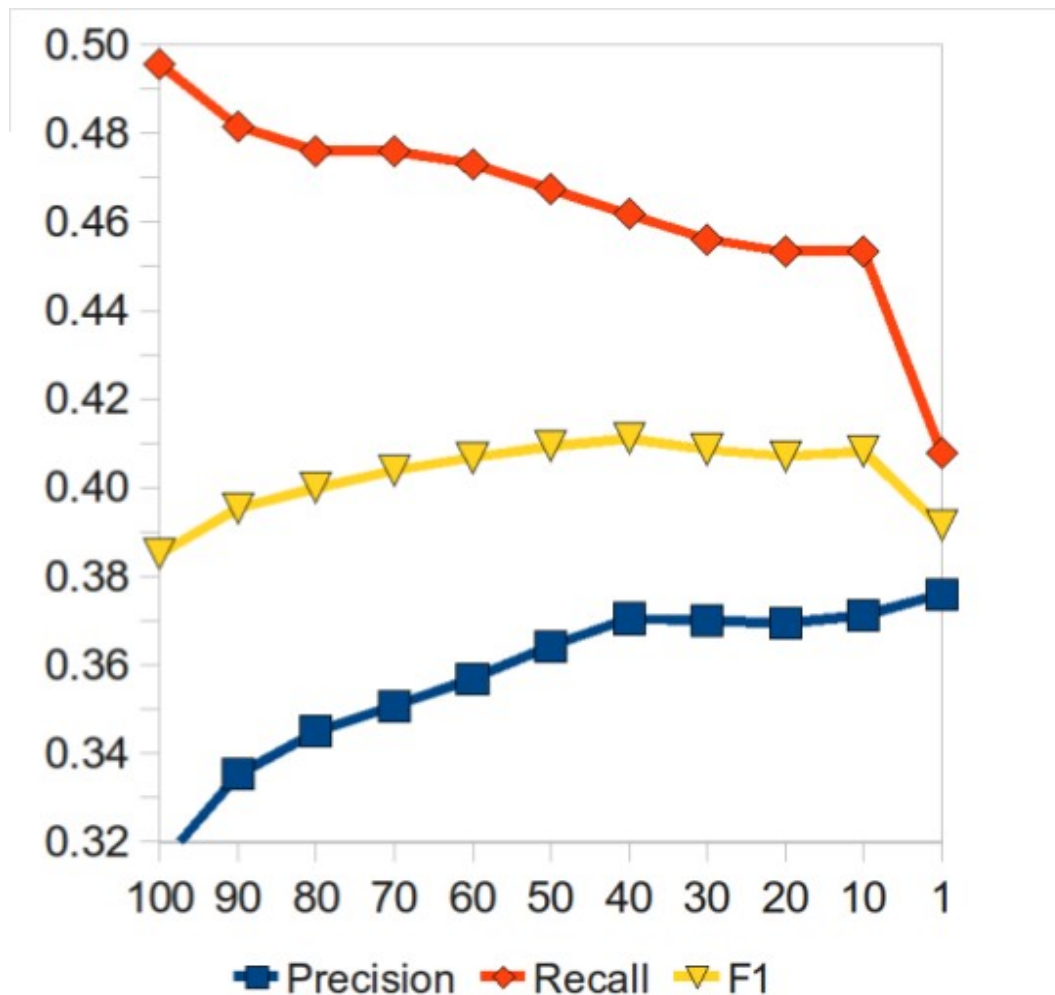


Figure 1: Results when keeping the top $N\%$ word senses according to the WSD scores.

Kybot output evaluation

- **Best performing** profiles

	# profiles	# triplets	# in scope	# correct	P.	R.	F1
WSD 60%	129	164	1226	434	0.38	0.46	0.42
WSD 60% & profiles 1%	104	912	332	147	0.44	0.42	0.43
WSD 60% & profiles 5%	103	775	312	147	0.47	0.42	0.44
WSD 60% & profiles 10%	103	775	312	147	0.47	0.42	0.44
WSD 60% & profiles 25%	93	693	284	141	0.50	0.40	0.44
WSD 60% & profiles 50%	76	523	219	115	0.53	0.33	0.40
WSD 60% & profiles 75%	22	119	46	32	0.70	0.09	0.16

Table 9: Generic processing with WSD threshold of 60% and using best performing profiles.

- Precision 50% vs. 39% manual WSD!
- Improve parsing (and Kybots) instead of better WSD

Kybot output evaluation

- **Generic vs. Domain**

	# triplets	# in scope	# correct	P.	R.	F1
generic processing with profiles	1816	548	174	0.32	0.49	0.39
domain processing with profiles	1528	509	156	0.31	0.44	0.36
domain processing with cterms	50	9	4	0.44	0.01	0.02
domain processing with profiles and cterms	1578	518	160	0.31	0.45	0.37

Table 10: Domain processing using the complex term heuristics.

- Generic processing better than domain !
 - Generic: 19,037 links to the ontology in the doc.
 - Domain: 16,953 links to the ontology in the doc.

Kybot output evaluation

- “airborne contaminant” **dw-eng-30-258-n**
 - sc_hasState Kyoto#airborne
 - sc_partOf DomainKyoto2#air
 - sc_subClassOf Kyoto#material__stuff-eng-3.0-14580897-n
- “contaminant” **eng-30-14821984-n**
 - sc_domainOf DOLCE-Lite.owl#amount-of-matter
 - sc_participantOf Kyoto#change__alter__modify-eng-3.0-00126264-v
 - sc_participantOf Kyoto#contamination__pollution-eng-3.0-00276987-n
 - sc_playRole Kyoto#done-by
 - sc_playRole Kyoto#use-of
 - sc_subClassOf Kyoto#material__stuff-eng-3.0-14580897-n
- “airborne” **eng-30-01522895-a**
 - sc_qualityOf Kyoto#quality-eng-3.0-04723816-n

Kybot output evaluation

- “suburban runoff” **dw-eng-30-221-n**
 - sc_hasCoParticipant Kyoto#city__metropolis-eng-3.0-08524735-n
 - sc_playCoRole Kyoto#has-source
 - sc_subClassOf Kyoto#happening__natural_event-eng-3.0-07283608-n
- “runoff” **eng-30-07407272-n**
 - sc_domainOf Kyoto#water-eng-3.0-07935504-n
 - sc_participantOf Kyoto#flow
 - sc_playRole Kyoto#done-by
 - sc_subClassOf Kyoto#commerce__mercantilism-eng-3.0-01090446-n
 - sc_subClassOf Kyoto#happening__natural_event-eng-3.0-07283608-n
 - sc_subClassOf Kyoto#move-eng-3.0-01831531-v
- “suburban” **eng-30-02804590-a**
 - sc_qualityOf Kyoto#district__territory__dominion-eng-3.0-08552138-n
 - sc_subClassOf Kyoto#quality-eng-3.0-04723816-n



Open competition

- Create **gold standard**
 - Choose three documents:
<http://www.thedailygreen.com>
 - Manually annotate events/roles
 - Convert events/roles to triplets
 - **KafAnnotator**
 - 2,003 tokens
 - Annotate 256 triplets
- Run Kybot profiles and measure precision/recall.

Results Open Competition

Overall results for all 3 files	# triplets	# in scope	# relations	# correct	R.	P.	F1
GS	256	256	253				
Baseline	14902	1815	1815	50	0.20	0.03	0.05
AST	15	8	8	3	0.01	0.38	0.02
KAIST	165	62	60	34	0.13	0.57	0.22
KYOTO	3461	964	192	58	0.23	0.30	0.26

Table 12: Results of the open competition.

Overall results ignoring the relations	# triplets	# in scope	# relations	# correct	R.	P.	F1
GS	256	256	253				
Baseline	50	1815	1815	155	0.61	0.09	0.15
AST	3	8	8	7	0.03	0.88	0.05
KAIST	34	62	60	47	0.18	0.78	0.30
KYOTO	58	964	192	85	0.33	0.44	0.38

Table 14: Results of the open competition, ignoring the relation

Transferring Kybots to another language

- Estuary database (**Dutch**)
- 93 documents, 42,697 words.
- 65 Dutch profiles adapted from English (**half a day work**)
- 4,095 events
- 6,862 roles
- 8,118 date expressions, 82 unique dates
- 5,928 place expressions, 60 unique GeoNames places
- 3,302 countries, 9 unique GeoNames countries

Transferring Kybots to another domain

- Medical protocols on the **treatment of breast cancer**
- 7 PDF documents, 110,501 word tokens
- No domain adaptation

- 8,416 events
- 15,984 roles
- Examples:
 - "disease includes tumour" ...
 - "axilla contains a sentinel lymph node"
 - "a side effect risk is part of the treatment"

Main results so far

- **KAF**
- **Generic Knowledge Architecture**
 - 3-layered model
 - Kyoto ontology
 - Formal mapping to **all** wordnets synsets
- **Robust Ontology-based IE**
 - Kybot, mining module
 - Off-line reasoning
 - Portable to other languages
 - Portable to other domains
- **Evaluation framework**

Current and future plans

- Chunk level queries
 - Search for a term and then a chunk whose head is ...
 - Inter-chunk searches
 - Search for a term and then, in the same chunk, another one which ...
- Dependency queries
- Layer-2 Kybots
 - Amalgamate events from several documents and languages
- Creating Kybots semi or fully automatically
 - Mining by example
 - Machine learning / Active Learning



KYOTO (ICT-211423) Intelligent Content and Semantics
Knowledge **Y**ielding **O**ntologies for **T**ransition-Based **O**rganization
<http://www.kyoto-project.eu/>

Event and Fact Mining

German Rigau
IXA group, UPV/EHU

Final Review
April 8th, 2011, Berlin, Germany



ICT-211423