A full knowledge cycle for semantic interoperability

Wauter Bosma & Piek Vossen
(VU University Amsterdam)
Eneko Agirre & Aitor Soroa & German Rigau
(Basque Country University)

KYOTO (ICT-211423) EU-FP7 ICT Program
http://www.kyoto-project.eu

ISA-5 and ISO/TC 37/SC4/WG 1-4 Meeting
15-17 January 2010, Hong Kong
Semantic interoperability

- Representation linear textual structure and composition
- Anchoring to conceptual and formal semantic implications
- Representation of factual knowledge and information embedded in text
- Representation of communicative function
- Representation of epistemic, deontic and interpersonal relations
- Compatible across languages and social groups
KYOTO – some statistics

- March 2008 – March 2011
- 8 countries (The Netherlands, Italy, Germany, Spain, Taiwan, Japan, Czech Republic)
- 12 sites
  - Universities & research institutes: VUA, CNR-ILC, CNR-IIT, BBAW, EHU, AS, NICT, Masaryk
  - Companies: Synthema, Irion
  - User organizations: ECNC, WWF
- 7 languages (English, Italian, Japanese, Dutch, Spanish, Basque, Chinese)
KYOTO – Overall architecture

1. Wordnets & Ontology
   - Multilingual Knowledge Base
   - Wiki Editor

2. Semantic & Syntactic representation
   - Kyoto Annotation Format

Linguistic Processor

Kybot
- Fact Extractor
- Term Extractor

Tybot
- Term Base
- Fact Base
KYOTO – Overall architecture

- Tokenization (language-sp)
- Parsing (language-sp)
- Word-sense-disambiguation
- Named entity recognition
- Ontological tagging
KYOTO – Overall architecture

- Linear text tokens
- Conceptual tags and relations
- Conceptual anchoring

Kyoto Annotation Format

Linguistic Processor

Wordnets & Ontology

Multilingual Knowledge Base

Fact Base

Term Base

Fact Extractor

Term Extractor

Kybot

Tybot
KYOTO – Overall architecture

- Non-linear
- Generic conceptual types & relations
- Conceptual anchoring

- Linear text tokens
- Conceptual tags and relations
- Conceptual anchoring

- Non-linear
- Concept instances and relations
- Conceptual anchoring

Linguistic Processor

Semantic & Syntactic representation
Kyoto Annotation Format

Multilingual Knowledge Base

Wordnets & Ontology

Fact Base

Fact Extractor

Kybot

Term Extractor

Term Base

Tybot

Kyoto Annotation Format
Requirements for semantic annotation in KYOTO

- Interoperability across languages and cultures
  - Language-neutral annotation of text, concepts and facts
  - One format for all languages
  - One fund of concepts for all languages
- Interoperability across linguistic processors
  - Specialized processors for specific tasks
  - System should work with new (unknown) languages
- Flexibility and extendibility, as requirements for applications may change over time
How to achieve semantic interoperability?

- Global wordnet grid model for anchoring vocabulary to formal concepts
- Kyoto Annotation Format based on LAF (Ide and Romary 2003) for interoperable and extendable annotation:
  - Linear text structures
  - Non-linear factual knowledge and information
- Reinterpretation of current ISO proposals
Global Wordnet Grid Model

**English Wordnet in WN-LMF**

- **bird_1_N**
  - **rigid**
- **migration_bird_1_N**
  - **non-rigid**
- **hyponym**
  - **duck_1_N**
  - **rigid**
- **migration_4_N**
- **migrate_1_V**
  - **sc_equivalentOf migration**

**KYOTO Ontology in OWL-DL**

(Extension of DOLCE LT)

- **perdurant**
- **endurant**
  - **role**
  - **object**
- **change-of-location**
- **done-by**
  - **has-path**
  - **has-destination**
  - **has-source**
- **migration**
  - **has**
  - **has**
  - **some**

**Dutch Wordnet**
**Spanish Wordnet**
**Basque Wordnet**
**Italian Wordnet**
**Japanese Wordnet**
**Chinese Wordnet**

...
Wordnet to ontology mappings

{create, produce, make}Verb, English
- sc_equivalenceOf construction

{artifact, artefact}Noun, English
- sc_domainOf physical_object
- sc_playRole result-existence
- sc_participateIn construction

{kunststof}Noun, Dutch // lit. artifact substance
- sc_domainOf amount_of_matter
- sc_playRole result-existence
- sc_participateIn construction

{meat}Noun, English
- sc_domainOf cow, sheep, pig
- sc_playRole patient

{名, 肉, 食物, 餐}Noun, Chinese
- sc_domainOf animal
- sc_playRole patient
- sc_playRole eat

{غذاء, لحم, طعام}Noun, Arabic
- sc_domainOf cow, sheep
- sc_playRole patient
- sc_playRole eat
Annotation consists of layers stacked on top of each other.

Layers are used to generate more sophisticated layers:

- **Morpho-syntactic layers** – language specific parsing
- **Level-1** semantic layers – named entities, events, etc.
- **Level-2** semantic layers – facts

Layers refer to items in lower level layers.

KAF is LAF-compliant.
Morpho-syntactic layers

- **Text**: tokenization, sentences, paragraphs, with reference to the source
- **Terms** [Text]: words and multi-words, includes parts-of-speech, declension information, etc.
- **Dependencies** [Terms]: dependency relations between terms
- **Chunks** [Terms]: constituents & phrases
Semantic layers

- Level-1 layers for **linear annotation**: tagging text elements (expressions of time, events, quantities, locations, etc.)

- Level-2 layers for **generic annotation**: extracted facts (with pointers to evidence in the text) – possibly **multiple** sources of evidence

- Linear vs. Generic ↔ Information vs. Knowledge
<kaf xml:lang="en">
  <kafHeader>...</kafHeader>
  layer 1...
  layer 2...
  ...
  ...
  layer N...
</kaf>
<kaf>
  <text>
    <wf wid="w1" page="1" sent="1" para="1"
      fileoffset="0,3">two</wf>
    <wf wid="w2" page="1" sent="1" para="1"
      fileoffset="4,7">per</wf>
    <wf wid="w3" page="1" sent="1" para="1"
      fileoffset="8,12">cent</wf>
  </text>
  <terms>
    <term tid="t1" type="open" lemma="two" pos="G">
      <span id="w1">!</span><!-- refers to "two" (w1) -->
    </term>
    <term tid="t2" type="open" lemma="per cent" pos="N">
      <span id="w2">html id="w2"></span><span id="w3"></span>
    </term>
  </terms>
</kaf>
Morpho-syntactic annotation: deps and chunks

<kaf>
  <text>...</text><!-- defines w1, w2, w3 -->
  <terms>...</terms><!-- defines t1, t2 -->

  <deps>
  <!-- dependency: "two" (t1) → "per cent" (t2) -->
  <dep from="t1" to="t2" rfunc="mod"/>
  </deps>

  <chunks>
  <!-- two per cent -->
  <chunk cid="c1" head="t2" phrase="NP">
    <span id="t1"/> <!-- refers to term: "two" -->
    <span id="t2"/> <!-- refers to term: "per cent" -->
  </chunk>
  </chunks>

</kaf>
<terms>
  <term tid="t1" type="open" lemma="bird" pos="N">
    <span id="w1"/>
  </term>
  <externalReferences>
    <!-- inserted by word-sense-disambiguation -->
    <externalRef resource="wn30g" ref="eng-30-01855672-n" conf="0.38"/>
    <externalRef resource="wn30g" ref="eng-30-10157744-n" conf="0.31"/>
    <externalRef resource="wn30g" ref="eng-30-07646821-n" conf="0.30"/>
  </externalReferences>
</terms>
<!-- Migration birds in the Humber Estuary -->
<term lemma="migration bird">
<externalReference>
<!-- Tagging terms with ontological implications based on wordnet mappings -->
<externalRef resource="ontology" relation="sc_domainOf" reference="bird"/>
<externalRef resource="ontology" relation="sc_participantOf" reference="migration"/>
<externalRef resource="ontology" relation="sc_playRole" reference="done-by"/>
<externalRef resource="ontology" relation="implied" reference="done-by" some="physical-plurality"/>
<externalRef resource="ontology" relation="implied" reference="has-destination" some="particular"/>
<externalRef resource="ontology" relation="implied" reference="has-source" some="particular"/>
<externalRef resource="ontology" relation="implied" reference="has-path" some="particular"/>
</externalReference>
</term>

<term lemma="in" pos="P"/>
<term lemma = "Humber Estuary"><externalRef resource="ontology" reference="locative-role"/></term>
Linear semantic annotation

<!-- Bird migration in the Humber Estuary -->
<term lemma="bird">
<externalReference><!-- Tagging terms with ontological implications based on wordnet mappings -->
<externalRef resource="ontology" relation="sc_equivalentOf" reference="bird"/>
</externalReference>
</term>

<term lemma="migration">
<externalReference><!-- Tagging terms with ontological implications based on wordnet mappings -->
<externalRef resource="ontology" relation="sc_equivalentOf" reference="migration"/>
<externalRef resource="ontology" relation="implied" reference="done-by" some="physical-plurality"/>
<externalRef resource="ontology" relation="implied" reference="has-destination" some="particular"/>
<externalRef resource="ontology" relation="implied" reference="has-source" some="particular"/>
<externalRef resource="ontology" relation="implied" reference="has-path" some="particular"/>
</externalReference>
</term>

<term lemma="in"/>
<term lemma = "Humber Estuary"><externalRef resource="ontology” reference="locative-role"/>
Linear semantic annotation

<!-- Birds that migrate to the Humber Estuary -->

<term lemma="bird" pos="N">  
<externalReference> <!-- Tagging terms with ontological implications based on wordnet mappings -->
  <externalRef resource="ontology" relation="sc_equivalentOf" reference="bird"/>
</externalReference>
</term>

<term lemma="migrate" pos="V">  
<externalReference> <!-- Tagging terms with ontological implications based on wordnet mappings -->
  <externalRef resource="ontology" relation="sc_equivalentOf" reference="migration"/>
  <externalRef resource="ontology" relation="implied" reference="done-by" some="physical-plurality"/>
  <externalRef resource="ontology" relation="implied" reference="has-destination" some="particular"/>
  <externalRef resource="ontology" relation="implied" reference="has-source" some="particular"/>
  <externalRef resource="ontology" relation="implied" reference="has-path" some="particular"/>
</externalReference>
</term>

<term lemma="to"/>

<term lemma = "Humber Estuary"><externalRef resource="ontology" reference="locative-role"/></term>
Adding layers of semantic annotation

- Kybot (Knowledge yielding robot)
- Kybot profiles, which are XML files with:
  - Regular expressions on different KAF layers
    - Variables for elements with structural & semantic properties
    - Relations between elements (e.g. Order)
  - Output pattern in any format
- Any number of profiles can be loaded and applied to document annotated in KAF
- Kybot profiles can be stacked, i.e. take the output of another Kybot as input
Examples of Kybot profile

<?xml version="1.0" encoding="utf-8"?>
<Kybot id="quantity-change001">
<variables>
    <var name="x" type="term.externalReference" relation="sc_equivalentOf" reference="migration"/>
    <var name="x" type="term.externalReference" reference="done-by"/>
    <var name="y" type="term.externalReference" relation="sc_subclassOf" reference="bird"/>
    <var name="p" type="term" lemma="from" pos="P"/>
    <var name="o" type="term.externalReference" relation="sc_subclassOf" reference="location"/>
</variables>
<!-- migration of birds from the Humber Estuary -->
<relations>
    <root span="x"/> <rel span="y" pivot="x" direction="following" immediate="true" />
        <rel span="o" pivot="p" direction="following" immediate="true" />
</relations>
<events>        <event eid='e1' target tid='$x/@tid' lemma='$x@lemma' pos='$x@pos'/> </events>
<roles>           <role rid='r1' target='$y@tid' event='$x/@tid' rtype='done-by'/>
    <role rid='r2' target='$o@tid' event='$x/@tid' rtype='has-source'/></roles>
</Kybot>
<events> <!-- event layer added to KAF by KYBOT profiles -->
  <event eid="e1" target="c2" lemma="teach" pos="V" tense="PAST"
    aspect="NONE" polarity="POS"/>
  <event eid="e2" target="c20" lemma="migrate" pos="V" tense="PRESENT"
    aspect="NONE" polarity="POS"/>
  <role rid="r1" event="e1" target="c1" rtype="done-by"/>
  <role rid="r2" event="e1" target="c3" rtype="patient"/>
  <role rid="r3" event="e1" target="c9" rtype="has-instrument"/>
  <role rid="r3" event="e2" target="c20" rtype="done-by"/>
  <role rid="r4" event="e2" target="c22" rtype="has-source"/>
  <role rid="r5" event="e2" target="c24" rtype="has-target"/>
</events>
Linear semantic annotation

```xml
<timexs> <!-- time objects added to KAF by KYBOT profiles -->

<!-- 1970 -->
<timex3 texid="timex1" type="DATE" value="1970">
   <span><target id="c7"/></span>
</timex3>

<!-- 2003 -->
<timex3 texid="timex2" type="DATE" value="2003">
   <span><target id="c9"/></span>
</timex3>

<!-- between 1970 and 2003 -->
<timex3 texid="timex3" type="DURATION" value="P33Y"
   beginPoint="timex1" endPoint="timex2"
   temporalFunction="true"/>

</timexs>
```
Non-linear annotation of facts

- Abstract from the linear order but refer back
- Amalgamate elements from multiple places in text, possibly across documents and across languages
- Split facts on the basis of coordination and quantification in language
- Match linear references of events and relations conceptually
- Limit events and relations within a time frame and region
Non-linear annotation of facts

- List of all unique entities occurring in the text or a document collection
  - Named entities
  - Conceptual entities
- List of all events and relations expressed in the text or document collection:
  - Event type
  - Roles → pointing to entities or text references
  - Time → pointing to entities
  - Place → pointing to entities
<location x4 id="l14">
  <span doc="767"> <target tid="t1873"/><target tid="t1271"/> </span>
  <span doc="897"> <target tid="t1471"/> </span>
  <externalReferences>
    <externalRef resource="GeoNames" ref="2635503"/>
  </externalReferences>
  <geoInfo>
    <place name="River Trent" countryCode="GB" countryName="United Kingdom" latitude="53.7" longitude="-0.7" fname="stream" timezone="Europe/London"/>
  </geoInfo>
</location>
Splitting and accumulating facts

DOC:6254 The Humber Estuary Low Tide Count Programme 2003-2004, (....) the recent recovery of the pinkfooted goose, avocet and black-tailed godwit populations.

DOC:212 Pinkfooted goose recovered as a result of better feeding conditions.

<fact fid="f1" type="recover"> <!-- recovery, recovered -->
  <span doc="6254"> <target eid="e45"/></span><span doc="212"><target eid="e35"></span>
  <fact:role rtype="experiencer"> <!--pinkfooted goose ->
    <span doc="6254"<target cid="c2"></span> <span doc="212"<target cid="c34"></span>
  </fact:role>
  <fact:role rtype="caused-by"> <!--better feeding conditions -->
    <span doc="212"<target cid="c4"> <target cid="c54"> <target cid="c254"></span>
  </span>
  </fact:role>
<timex3 entid="t3"> <!-- 2003, 2004 -->
<locationx4 lid="l711"> <!-- Humber Estuary -->
</fact>

.... similar for avocet and for black-tailed godwit but better feeding conditions do not apply
KAF in KYOTO

- Word Sense Disambiguation adds sense annotation to the **term** layer of KAF
- Ontotagger inserts implied semantics into the **term** layer
- Tybots (term yielding robots) use KAF for **term extraction**
  - Uses the **term** layer and the **chunk** layer
- Kybots (knowledge yielding robots) use KAF for **fact extraction**
  - Kybot is configured to search for specific facts by defining a **kybot profile**
- All of the above are mostly **language-neutral** applications
ISO & W3C standards

- Wordnet-LMF is based on LMF, Kyoto ontology is implemented in OWL-DL
- Background vocabularies are in SKOS and acquired Terms in TMF
- KAF is inspired by: SynAF (dependency relations), MAF (morphological annotation), SemAF (time and events), LAF (generic linguistic annotation framework)
- SynAF, MAF and SemAF cannot be stacked
- LAF is a data model rather than a standard
- KAF is an instantiation of LAF with elements from SynAF, MAF and SemAF
Conclusion

- Separate the textual semantic encoding from the conceptual relations from the instantiation of relations to achieve semantic interoperability
- Global Wordnet Grid for conceptual anchoring of language
- Key features of KAF:
  - Layered annotation; extendible for new applications
  - Distributed and language neutral processing
  - Sharing & reusing resources
- KAF in KYOTO:
  - Three types of annotation: morphosyntactic, linear (level-1 semantic) and generic (level-2 semantic)
  - Used for 7 languages in several applications