KAF: a generic semantic annotation format

Wauter Bosma & Piek Vossen (VU University Amsterdam)
Aitor Soroa & German Rigau (Basque Country University)
Maurizio Tesconi & Andrea Marchetti (CNR-IIT, Pisa)
Carlo Aliprandi (Synthema, Pisa)
Monica Monachini (CNR-ILC, Pisa)

KYOTO
EU-FP7 ICT Program
KYOTO – overview

- A system for defining and sharing meaning in a domain
  - Domain wordnet (linked to generic wordnet)
  - Ontology (linked to wordnet)
  - Fact profiles
- Semantic interoperability
- Knowledge is maintained by end-users
- System can be used for extracting factual data from documents
- Cross-language; cross-culture
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 – knowledge cycle

Wiki environments
Bridging cultures

Documents

Websites
PDF documents

Community

Terminology

Extracted facts
Accumulated knowledge

Knowledge

Extracted domain terms
Ontologies
Requirements for semantic annotation in KYOTO

- Interoperability across languages and cultures
  - Language-neutral annotation
  - One format 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
The KYOTO way

- **KAF**: KYOTO/Knowledge Annotation Format
- 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"/> <!-- refers to "two" (w1) -->
    </term>
    <term tid="t2" type="open" lemma="per cent" pos="N">
      <span id="w2"/><span id="w3"/>
    </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>
<timexs>

<!-- 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>
<entities>
  <ent eid="e1"> <!-- change -->
    <spans>
      <span><target doc="134" id="c7"/></span>
      <span><target doc="134" id="c34"/></span>
      <span><target doc="14" id="c13"/></span>
    </spans>
  </ent>
  <ent eid="e300"> <!-- change -->
    <spans>
      <span><target doc="134" id="c13"/></span>
      <span><target doc="4" id="c3"/></span>
    </spans>
  </ent>
</entities>
Generic annotation

<facts>
<!-- Source: between 1970 and 2003, tropical Species [...] Temperate species populations have shown little overall change. -->
<!-- Fact: change(temperate species populations, little, 1970–2003) -->
<fact fid="f1">
<!-- change -->
<process eid="e1"/>
<!-- little -->
<quantity qid="q1"/>
<!-- between 1970 and 2003 -->
<timex3 texid="timex3"/>
<!-- temperate species populations -->
<arg tid="c1" role="patient"/>
</fact>
</facts>
KAF in KYOTO

- Word Sense Disambiguation adds sense annotation to the **terms** layer of KAF
- Tybots (term yielding robots) use KAF for **term extraction**
  - Uses the **terms** layer and the **chunks** layer
- Kybots (knowledge yielding robots) use KAF for **fact extraction**
  - Kybot is configured to search for specific facts by defining a **kybot profile**
- Wikyoto allows domain experts to define **kybot profiles** and to build a **domain wordnet** from Tybot terms, linked to a shared ontology
- All of the above are **language-neutral**
KAF and ISO standards

- **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

- Key features of KAF:
  - Layered annotation; extendible for new applications
  - Distributed processing
  - 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

- KAF manual: www.kyoto-project.eu (under system architecture and demos, data formats)