Frame-based Ontology Population from Text with PIKES

Francesco Corcoglioniti
Fondazione Bruno Kessler – Trento, Italy
corcoglio@fbk.eu

joint work with Marco Rospocher and Alessio Palmero Aprosio

http://pikes.fbk.eu/
Knowledge Extraction (KE) from text
- cross-disciplinary field (SW, NLP & KR) allowing bridging Web & Semantic Web
- a.k.a. **Ontology Learning & Population** [Cimiano, 2006]
  - Ontology Learning: extraction of terminological knowledge (TBox)
  - Ontology Population: extraction of assertional knowledge (Abox)
- knowledge items: instances, binary relations, **semantic frames**

**Semantic Frames**
- n-ary relations, events & structured entities reified as **ontological instances**
- linked by **role properties** to participant instances
- schema defined by **frame-based ontologies**, e.g., FrameBase, ESO

![Diagram](https://example.com/diagram.png)
Frame-based Ontology Population

Why?
- **fixed Tbox** favors using extracted knowledge (compare with Open IE)
- **expressivity** of frames better than binary relations
- **Semantic Role Labeling (SRL)** fits well (and also other NLP tasks)

Example: “G. W. Bush and Bono are very strong supporters of the fight of HIV in Africa.”

```
predicate: supporter.01
```

```
argument: argm-mnr
argument: arg0
argument: arg1
```

```
:fight_event_1
frb:fe-hostile_encounter
```

```
:support_event_1
frb:fe-taking_sides
```

```
dbpedia:Bono
dbyago:Person10007846
dbpedia:Bush
dbyago:Person10007846
dbpedia:HIV
owl:Thing
dbpedia:Africa
dbyago:Location100027167
```

```
attr:very-1r_strong-1a
ks:Attribute
```

```
frb:fe-taking_sides-cognizer
```

```
frb:fe-taking_sides-degree
```

```
frb:fe-hostile_encounter-side_2
frb:fe-hostile_encounter-place
frb:fe-taking_sides-issue
```
Introducing PIKES* – a tool for Frame-based Ontology Population from English text
- extracting
  • semantic frames typed w.r.t. FrameBase
  • instances typed w.r.t. YAGO and disambiguated w.r.t. DBpedia
- representing all contents in RDF + named graph
- based on a 2-phase approach
- open source

(*) PIKES Is a Knowledge Extraction Suite
A linguistically annotated piece of text about something of interest

:resource1 a ks:Resource; 
dct:created "2016-04-06"; 
nif:isString "G. W. Bush and Bono are very strong supporters of the fight of HIV in Africa."

:mention1 a ks:FrameMention; 
ks:mentionOf :resource1; 
nif:beginIndex 36; nif:endIndex 46; 
nif:anchorOf "supporters"; 
ks:synset wn30:n-10677713 
ks:predicate pmo:nb10_support.01; 
ks:role pmo:nb10_support.01_arg01; 
ks:expresses :graph1; 
ks:denotes :supporters_entity; 
ks:implies :support_event_1.

:graph1 { 
:supporters_entity 
a dbyago:Supporter110677713. 
:support_event_1 a 
a frb:frame-taking_sides; 
frb:fe-taking_sides-cognizer 
:supporters_entity. 
}

based on: Corcoglioniti et al. KnowledgeStore: a storage framework for interlinking unstructured and structured knowledge. IJSWIS 2015
Data Model (2)

Complete model:

- **ks:Resource**
  - dct:title
  - dct:creator
  - dct:created

- **ks:Instance**
  - rdf:type
  - rdfs:label
  - foaf:name

- **ks:Expression**
  - contains triples about
  - owl:sameAs
  - rdfs:seeAlso
  - ks:include

- **ks:Frame**
  - ks:FrameMention
    - ks:frame
  - ks:Predicate
  - ks:Argument

- **ks:Time**
  - ks:TimeMention
    - nif:beginIndex
    - nif:endIndex
    - nif:anchorOf
  - ks:synset
  - ks:linkedTo
  - ks:argument
  - ks:coreferential
  - ks:coreferentialConjunct

- **ks:Attribute**
  - ks:AttributeMention
    - ks:norm.Value
    - ks:normalizedValue

- **ks:Coreference**
  - ks:CoreferenceMention
    - ks:mentionOf

- **ks:Mention**
  - ks:mentionOf
  - ks:denotes / ks:implies

- **ks:Relation**
  - ks:RelationMention
    - ks:coreferential

- **ks:Entity**

- **ks:Resource**
  - dct:title
  - dct:creator
  - dct:created

- **nif:** <http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#>
- **ks:** <http://dkm.fbk.eu/ontologies/knowledgestore#>
- **foaf:** <http://xmlns.com/foaf/0.1/>
2-Phase Approach

Resource layer

G. W. Bush and Bono are very strong supporters of the fight of HIV in Africa.

Mention layer

Phase 1
Linguistic Feature Extraction

Instance layer

Phase 2
Knowledge Distillation

dbpedia:Bono
attr:very-1r_strong-1a
:support

frb:fe-taking_sides-degree

G. W. Bush and Bono [...] supporters

frb:fe-hostile_encounter-place

frb:fe-taking_sides-cognizer

G. W. Bush

frb:fe-hostile_encounter-degree

Bono

ks:coreferential

very strong

supporters

very strong supporters

supporters of [...] fight

fight of HIV

fight [...] in Africa

HIV

Africa

dbpedia:Africa

frb:fe-hostile_encounter-issue

frb:fe-taking_sides-issue

frb:fe-taking_degrees
Linguistic Feature Extraction

① apply several NLP tasks to input text
② map their outputs to mentions

<table>
<thead>
<tr>
<th>NLP Task</th>
<th>Type of mention</th>
<th>Instance</th>
<th>Name</th>
<th>Time</th>
<th>Attribute</th>
<th>Frame</th>
<th>Participation</th>
<th>Coreference</th>
</tr>
</thead>
<tbody>
<tr>
<td>part-of-speech tagging</td>
<td>POS</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>named entity recognition &amp; classification</td>
<td>NERC</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>temporal expression recognition &amp; norm.</td>
<td>TERN</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>entity linking</td>
<td>EL</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>word sense disambiguation</td>
<td>WSD</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>semantic role labeling</td>
<td>SRL</td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>coreference resolution</td>
<td>COREF</td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dependency parsing</td>
<td>DP</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Example:

"fight of HIV"

via NERC, EL

Extracted RDF mention graph

via SRL

via SRL, DP

```
<..#char=63,66> a :NameMention ;
nif:anchorOf "HIV" ;
:nercType :MISC ;
:linkedTo dbpedia:HIV .
```

```
<..#char=54,59> a :FrameMention ;
nif:anchorOf "fight"
:predicate pm:nb10-fight.01 .
```

```
<..#char=54,66> a :ParticipationMention;
nif:anchorOf "fight […] HIV"
:frame <..#char=54,59> ;
:argument <..#char=63,66> ;
:role pmo:nb10-fight.01-arg1 .
```
Knowledge Distillation

① Rule-based conversion

Mention layer
- mention graph
- mention attributes

Background knowledge (BK)
- mappings (e.g., PreMOn)
- further metadata (e.g., argument nominalization)

Instance layer
- instances (entities, semantic frames)
- triples (participation, owl:sameAs, …)
- named graphs
- links to mentions

Rules

Example:

body: mention with predicate annotation + predicate triggers argument nominalization (bk)
head: create semantic frame & argument instances

→ rules can be expressed in SPARQL

② Post-processing
- inference and owl:sameAs smushing (i.e., merge owl:sameAs instances)
- named graph merging in presence of same metadata / linked mentions
Knowledge Distillation (2)

Mapping rules example:

**Mention layer**

```
:mention1 a ks:FrameMention;
  nif:anchorOf "supporters";
  ks:synset wn30:n-10677713;
  ks:predicate pmo:nb10_support.01;
  ks:role pmo:nb10_support.01_arg01;
```

**Instance layer**

```
:g1 {  :e1 a dbyago:Supporter110677713.
     :ev1 a frb:frame-taking_sides;
     frb:fe-taking_sides-cognizer :e1. }

:mention1 ks:expresses :g1;
  ks:denotes :e1; ks:implies :ev1.
```

**Background knowledge**

```
pmo:nb10_support.01
  a ks:ArgumentNominalization.
```

**Graph Query**

```
INSERT { ?m ks:denotes ?i; ks:implies ?if; ks:expresses ?g.}
GRAPH ?g { ?i a ks:Instance. ?if a ks:Frame } }
  ?s a ks:ArgumentNominalization. }
BIND (ks:mint(?m) AS ?g) BIND (ks:mint(?a, ?m) AS ?i)
BIND (ks:mint(concat(?a, "_pred"), ?m) AS ?if)
```
Implementation

PIKES
- Java 1.8 on Linux / Mac OS X
- open source (GPL)
- Maven project on GitHub
  https://github.com/dkmfbk/pikes

Integrated dependencies
- Stanford CoreNLP
- Mate-tools
- Semafor
- RDFpro

External dependencies
- Dbpedia Spotlight
- UKB

→ need separate install
PIKES UI for:

“G.W. Bush and Bono are very strong supporters of the fight of HIV in Africa. Their March 2002 meeting resulted in a 5 billion dollar aid.”

http://pikes.fbk.eu/
1\textsuperscript{st} Evaluation: Throughput

Setup
- **109K Wikipedia-like pages** (source: Simple English Wikipedia)
- **16 PIKES instances** processing pages in parallel
- **24 CPU cores** (dual Xeon E5-2430), 192GB ram, SSD

Results
- **32 hours** total processing time
- **358M triples** (2M resource, 283M mention, 72M instance layers)

<table>
<thead>
<tr>
<th>item type</th>
<th># items</th>
<th>throughput [item/h]</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>24 cores</td>
</tr>
<tr>
<td>Documents</td>
<td>109,242</td>
<td>3,450</td>
</tr>
<tr>
<td>Sentences</td>
<td>1,584,406</td>
<td>50,000</td>
</tr>
<tr>
<td>Tokens</td>
<td>23,877,597</td>
<td>753,000</td>
</tr>
</tbody>
</table>

(* estimated based on time using all cores)
2nd Evaluation: FrameBase Population

Gold standard
- 8 sentences (233 tokens) from [Gangemi, ESWC 2013]
- 166 triples (137 instances, 155 edges) manually annotated by 2 annotators

Three SRL configurations
- Mate-tools – PropBank, NomBank → FrameBase via our mappings
- Semafor – FrameNet → FrameBase via existing mappings
- Mate-tools + Semafor leveraging outputs combination in Mention layer

<table>
<thead>
<tr>
<th></th>
<th>Mate-tools</th>
<th>Semafor</th>
<th>Mate-tools + Semafor</th>
</tr>
</thead>
<tbody>
<tr>
<td>types</td>
<td>0.792</td>
<td>0.617</td>
<td>0.792</td>
</tr>
<tr>
<td>roles</td>
<td>0.633</td>
<td>0.594</td>
<td>0.704</td>
</tr>
<tr>
<td>triples</td>
<td>0.704</td>
<td>0.595</td>
<td>0.704</td>
</tr>
</tbody>
</table>

Precision: Mate-tools 0.792, Semafor 0.633, Mate-tools + Semafor 0.704
Recall: Mate-tools 0.617, Semafor 0.594, Mate-tools + Semafor 0.633
F1: Mate-tools 0.698, Semafor 0.435, Mate-tools + Semafor 0.526

with Mate-tools: low recall & F1 due to mappings
with both tools: + 0.061 F1 roles + 0.035 F1 all
3rd Evaluation: Comparison with FRED

- PIKES vs FRED [Presutti et al, EKAW 2012] on gold standard
- differently from PIKES, FRED is also able to extract TBox axioms
- comparison on VerbNet (VN) & FrameNet (FN) extracted triples
- PIKES nominal frames converted to FRED binary relation
  - e.g., “Iraqi official”: { :rel1 :arg :per1, dbpedia:Iraq } → { :per1 :rel dbpedia:Iraq }

<table>
<thead>
<tr>
<th>Category</th>
<th>Precision (PIKES)</th>
<th>Recall (FRED)</th>
<th>F1 (PIKES)</th>
<th>Recall (FRED)</th>
<th>F1 (PIKES)</th>
</tr>
</thead>
<tbody>
<tr>
<td>instances</td>
<td>0.911</td>
<td>0.93</td>
<td>0.971</td>
<td>0.869</td>
<td>0.94</td>
</tr>
<tr>
<td>edges</td>
<td>0.698</td>
<td>0.555</td>
<td>0.787</td>
<td>0.555</td>
<td>0.844</td>
</tr>
<tr>
<td>triples</td>
<td>0.543</td>
<td>0.416</td>
<td>0.636</td>
<td>0.471</td>
<td>+0.165 F1</td>
</tr>
<tr>
<td>types (VN)</td>
<td>0.667</td>
<td>0.593</td>
<td>0.667</td>
<td>0.516</td>
<td>0.552</td>
</tr>
<tr>
<td>types (FN)</td>
<td>0.731</td>
<td>0.55</td>
<td>0.526</td>
<td>0.478</td>
<td>0.615</td>
</tr>
<tr>
<td>roles (VN)</td>
<td>0.547</td>
<td>0.741</td>
<td>0.778</td>
<td>0.737</td>
<td>0.516</td>
</tr>
<tr>
<td>linking</td>
<td>0.7</td>
<td>0.615</td>
<td>0.444</td>
<td>0.545</td>
<td></td>
</tr>
<tr>
<td>coreference</td>
<td>0.357</td>
<td>0.857</td>
<td>0.4</td>
<td>0.333</td>
<td>0.345</td>
</tr>
</tbody>
</table>
Conclusions

PIKES is
- a tool for Frame-based Ontology Population from English text
- extracting events and complex relations (semantic frames)
- representing all contents in RDF + named graph
- featuring 2-phases: linguistic feature extraction + knowledge distillation

Benefits
- competitive with state of the art in terms of quality / throughput
- 2-phase decoupling allows tuning the two phases independently

Related/ongoing work
- PreMO - lemon extension for predicate models
- KE4IR - knowledge extraction for IR
- KnowledgeStore - store for PIKES data
- KEM - RDF/OWL model for knowledge extraction (ongoing work)
PreMO\textsuperscript{n} = \textbf{P}redicate \textbf{M}odel for \textbf{O}ntologies

[Corcoglioniti, Palmero Aprosio, Rospocher, Tonelli, LREC 2016] \hspace{1cm} \texttt{http://premon.fbk.eu/}

Linguistic Linked Data resource (grounded in Lemon) representing \textit{predicate models} and \textit{mapping} resources: PB, NB, VN, FN, Semlink

Homogeneously represents the \textit{semantic classes} (e.g., rolesets in NB and PB, verb classes in VN, frames in FN) and \textit{semantic roles}

Benefits:
- ease of \textit{access} and \textit{reuse} of predicate model data
- abstract \textit{commonalities}, keep \textit{peculiarities}
- automated \textit{reasoning} and \textit{SPARQL} querying
- \textit{interlinking} with third-party datasets

Availability: download / SPARQL endpoint / URI dereferencing
Frame-based Ontology Population from text with PIKES @ Ontology Summit 2017

FrameBase (FB)
8151 classes, 9632 props

PropBank 1.0
5576 / 5576 / 76241
517K triples

PropBank 1.7
6196 / 6181 / 145586
630K triples

PropBank 2.1.5
7671 / 8751 / 206701
926K triples

VerbNet 3.2
4402 / 484 / 1396
237K triples

FrameNet 1.5
9413 / 1018 / 9633
332K triples

FrameNet 1.6
10107 / 1204 / 11251
411K triples

WordNet 3.1 synsets
5248, –, –
(VN 3.2)

OntoNotes 5 groupings
3617, 1713, 1904
(SemLink 1.2.2c)

NomBank 1.0
5576 / 5576 / 76241
517K triples

596 concept. → 596 FB classes
491 roles → 252 FB props

Produce with PreMONitor
http://premon.fbk.eu/premonitor.html
KE4IR = Knowledge Extraction for Information Retrieval


PIKES analysis of query and documents to improve IR performances

Semantics considered (e.g. “astronomers influenced by Gauss”)
- **URIs**: dbpedia:Carl_Friedrich_Gauss
- **TYPE**: dbyago:Astronomer109818343, dbyago:GermanMathematicians
- **FRAME**: framebase:Subjective_influence
- **TIME**: dbo:dateOfBirth (1777), dbo:dateOfDeath (1855)

<table>
<thead>
<tr>
<th>Approach</th>
<th>Prec@1</th>
<th>Prec@5</th>
<th>Prec@10</th>
<th>NDCG</th>
<th>NDCG@10</th>
<th>MAP</th>
<th>MAP@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Textual</td>
<td>0.943</td>
<td>0.669</td>
<td>0.453</td>
<td>0.832</td>
<td>0.782</td>
<td>0.733</td>
<td>0.681</td>
</tr>
<tr>
<td>KE4IR</td>
<td>0.971</td>
<td>0.680</td>
<td>0.474</td>
<td>0.854</td>
<td>0.806</td>
<td>0.758</td>
<td>0.713</td>
</tr>
<tr>
<td>improvement</td>
<td>3.03%</td>
<td>1.71%</td>
<td>4.55%</td>
<td>2.64%</td>
<td>2.99%</td>
<td>3.50%</td>
<td>4.74%</td>
</tr>
</tbody>
</table>
KnowledgeStore

Scalable storage for text, NLP annotations and extracted knowledge


Any application
HTTP access, possibly exploiting SPARQL client libs.

Any Java application
KnowledgeStore Client

KnowledgeStore Client

KnowledgeStore Tools
populator, exporter, statistics collector, ...

Client-side

Server-side

KnowledgeStore Server
(single node)

CRUD endpoint

SPARQL endpoint

Web UI

Hadoop HDFS
(name & data nodes)

HBase
(mult. master & region nodes)

Virtuoso
(single node)

Representation

Resource

Mention

Instance

Scalable storage for text, NLP annotations and extracted knowledge

Revision of PIKES data model grounded on semiotics and NIF (ongoing work)

**Resource layer**

- `kem:CompositeResource`
- `kem:Resource`
- `kem:CompositeFragment`
- `kem:Fragment`
- `kem:Annotation`
- `kem:SemanticAnnotation`

**Mention layer**

- `kem:involves /
  kem:subject`
- `kem:isAbout /
  kem:refersTo`
- `contains RDF`
- `triples about`
- `kem:involvesSubjectOf`
- `kem:involvesReferentOf`
- `kem:hasAnnotation`
- `kem:hasResourceAnnotation`
- `kem:fragmentOf`
- `1..*`
- `2..*`

**Instance layer**

- `kem:Instance`
- `kem:Graph`
- `kem:conveys`
- `dbpedia:United_Nations`
- `rdf:type dbo:Organisation;`

---

**Indonesia Hit By Earthquake**

A United Nations assessment team was dispatched to the province after two quakes, measuring 7.6 and 7.4, struck west of Manokwari Jan. 4. At
Thank you!

Francesco Corcoglioniti
Fondazione Bruno Kessler – Trento, Italy
corcoglio@fbk.eu

joint work with Marco Rospocher
and Alessio Palmero Aprosio

http://pikes.fbk.eu/