Knowlywood: Mining Activity
Knowledge from Hollywood Narratives

Niket Tandon (MPI Informatics, Saarbruecken)
Gerard de Melo (IIIS, Tsinghua Univ)
Abir De (IIT Kharagpur)
Gerhard Weikum (MPI Informatics, Saarbruecken)
Legs, person, shoe, mountain, rope..
Rock climbing
Going up a mountain/ hill
Going up an elevation

Daytime, outdoor activity
What happens next?
<table>
<thead>
<tr>
<th>Rock climbing</th>
<th>Activity classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Going up a mountain/hill</td>
<td>Activity groupings</td>
</tr>
<tr>
<td>Going up an elevation</td>
<td>Activity hierarchy</td>
</tr>
<tr>
<td>Daytime, outdoor activity</td>
<td>Additional information</td>
</tr>
<tr>
<td>What happens next?</td>
<td>Temporal guidance</td>
</tr>
</tbody>
</table>
Go up an elevation

Previous activity

Get to village

Parent activity

{Climb up a mountain, Hike up a hill}

Next activity

Drink water

Participants: climber, boy, rope

Location: camp, forest, sea shore

Time: daylight, holiday

Visuals
## Activity commonsense: Related work

### Event mining

**Encyclopedic KBs:**
- Factual e.g. bornOn
- Entity oriented e.g. Person
- Many KBs: e.g. Freebase

<table>
<thead>
<tr>
<th>Alexander J. Hennold</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image" alt="Image of Alexander J. Hennold at the Tiff Film Festival" /></td>
</tr>
</tbody>
</table>

**Personal Information**
- **Born:** August 17, 1985 (age 29)
- **Education:** UC Berkeley (dropped out)
- **Occupation:** Professional rock climber

**Climbing career**
- **Type of climber:**
  - Free solo
  - Big wall
- **Highest grade:**
  - Redpoint: 5.14c (6+)
  - Bouldering: Y12 (8A+)
- **Known for:**
  - Big Wall Free Soloing
  - Speed record on The Nose of El Capitan
### Activity commonsense: Related work

<table>
<thead>
<tr>
<th>Event mining</th>
<th>Commonsense KB</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Encyclopedic KBs:</strong></td>
<td><strong>Cyc:</strong></td>
</tr>
<tr>
<td>Factual e.g. bornOn</td>
<td>Manual</td>
</tr>
<tr>
<td>Entity oriented e.g. Person</td>
<td>Limited size</td>
</tr>
<tr>
<td>Many KBs: e.g. Freebase</td>
<td>No focus on activities</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Commonsense KB</th>
<th><strong>ConceptNet:</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Crowdsourced</td>
</tr>
<tr>
<td></td>
<td>Limited size</td>
</tr>
<tr>
<td></td>
<td>No semantic activity frames</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Commonsense KB</th>
<th><strong>WebChild:</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No focus on activities</td>
</tr>
</tbody>
</table>

---

**Alexander J. Henmold**

- **Burn:** August 17, 1985 (age 30)
- **Education:** UC Berkeley (dropped out)
- **Occupation:** Professional rock climber
- **Type of climber:** Free solo, Big wall
- **Highest grade:** Redpoint 5.14c 6b+)
  - Bouldering: Y12 (8A+)
- **Known for:** Big Wall Free Soloing, Speed record on The Nose of El Capitan
# Activity commonsense: Related work

## Event mining

<table>
<thead>
<tr>
<th>Encyclopedic KBs:</th>
<th>Commonsense KB</th>
<th>This talk</th>
</tr>
</thead>
<tbody>
<tr>
<td>Factual e.g. bornOn</td>
<td>Cyc: Manual Limited size No focus on activities</td>
<td>Semantic Activity CSK KB construction</td>
</tr>
<tr>
<td>Entity oriented e.g. Person</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Many KBs: e.g. Freebase</td>
<td>ConceptNet: Crowdsourced Limited size No semantic activity frames</td>
<td></td>
</tr>
<tr>
<td></td>
<td>WebChild: No focus on activities</td>
<td></td>
</tr>
</tbody>
</table>
Activity commonsense is **hard**: 
- People hardly express the obvious: implicit and scarce 
- Spread across multiple modalities: text, image, videos 
- Non-factual: hence noisy
Hollywood narratives are easily available and meet the desiderata

**EXT. SMALL MOUNTAIN--DAY**

Wichita charges up the rockage of a small mountain-hill-type thing. The image repeats itself over and over--each time Wichita is more sweaty, gasping, sneering.

Wichita (V.O.)
The rules forbid anyone from the climbing the camp's mountain.

Contain events but not activity knowledge

May contain activities but varying granularity and no visuals. No clear scene boundaries.

align via subtitles with approximate dialogue similarity
State of the art WSD customized for phrases

Syntactic and semantic role semantics from VerbNet

the man

began to shoot

a video

man.1

man.2

shoot.1

shoot.4

video.1

NP

agent. animate

VP

shoot.vn.1

NP

patient. animate

NP

agent. animate

VP

shoot.vn.3

NP

patient. inanimate
State of the art WSD customized for phrases

The man

Began to shoot

A video

NP: agent. animate

VP: shoot.vn.1

NP: patient. animate

Output Frame

Agent: man.1

Action: shoot.4

Patient: video.1

Syntactic and semantic role semantics from VerbNet

Output

Frame

Agent

Action

Patient

Semantic Parsing

ILP for WSD + SRL

Graph Inference

Taxonomic + Statistical Priors

PSL Inference

Taxonomy Construction

Activity Synsets

Hierarchy

Knowlywood

Activity Knowledge Base
<table>
<thead>
<tr>
<th>IMS prior</th>
<th>WN prior</th>
<th>Word, VN match score</th>
<th>Selectional restriction score</th>
</tr>
</thead>
</table>

**maximize**

\[
\sum_{i, j} x_{ij} (\alpha \tau_{ij} + \beta_1 \theta_{ij} + \beta_2 \text{syn}_{ij} + \beta_3 \text{sem}_{ij})
\]

**subject to**

\[
\sum_{j \in S_V} x_{ij} \leq 1 \quad \forall i \in V
\]

\[
x_{ij} \leq x_{ij'} \quad \forall i \in V, j \in S_W, j \text{ mapped to } j' \in S_V
\]

\[
x_{i_0,j_0} \leq x_{ij} \quad \forall i_0 \in V, j \in S_V, x_{ij} \in \text{role-restr}(x_{i_0,j_0})
\]

\[
\sum_{j} x_{ij} \leq 1 \quad \forall i \notin V
\]

\[
x_{ij} \in \{0, 1\}
\]

\(x_{ij} = \text{binary decision var. for word } i, \text{ mapped to WN sense } j\)

**One VN sense per verb**

**WN, VN sense consistency**

**Selectional restr. constraints**

**binary decision**
Go up an elevation

Climb up a mountain
Participants: climber, rope
Location: camp, forest
Time: daylight

Hike up a hill
Participants: climber
Location: sea shore
Time: holiday

Drink water

Similarity:
\[
\frac{1}{2} \left( \frac{1}{1 + \text{dist}(v_i, v_j)} + \frac{1}{1 + \text{dist}(o_i, o_j)} \right) + \text{Attribute overlap}
\]

Hypernymy: WordNet hypernymy: \(v_i, v_j\) and \(o_i, o_j\) + Attribute hypernymy

Temporal: Generalized Sequence Pattern mining over statistics with gaps
\[#(\text{asynset}_1 \text{ precedes asynset}_2) / #(\text{asynset}_1) \#(\text{asynset}_2)\]
Probabilistic soft logic
- refining Typeof (T), Similar (S) and Prev (P) edges

1. Parents often inherit prev. (P) edges from their children:
   \[ P(a, b) \land T(a, a') \land T(b, b') \Rightarrow P(a', b'). \]

2. Similar activities are likely to share parent types
   \[ S(a, b) \land T(b, b_0) \Rightarrow T(a, b_0). \]

3. Likely mutual exclusion between edge types:
   \[ T(a, b) \land S(a, b) \Rightarrow \neg P(a, b). \]

4. Siblings are likely to be similar:
   \[ T(a, c) \land T(b, c) \Rightarrow S(a, b). \]

5. Similarity is often transitive:
   \[ S(a, b) \land S(b, c) \Rightarrow S(a, c). \]

6. Similarity is normally symmetric:
   \[ S(a, b) \Rightarrow S(b, a). \]
Climb up a mountain
- Participating Agent: climber, rope
- Location: camp, forest
- Time: daylight

Hike up a hill
- Participating Agent: climber
- Location: sea shore
- Time: holiday

Drink water

Tie the activity synsets

Break cycles

Resultant: DAG
Recap

- Defined a new problem of automatic acquisition of semantically refined frames.
- Proposed a joint method that needs no labeled data.
# Evaluation

<table>
<thead>
<tr>
<th>Knowlywood</th>
<th>Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scenes</td>
<td>1,708,782</td>
</tr>
<tr>
<td>Activity synsets</td>
<td>505,788</td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.85 ± 0.01</td>
</tr>
<tr>
<td>URL</td>
<td>bit.ly/knowlywood</td>
</tr>
</tbody>
</table>

#Scenes is aggregated counts over Moviescripts, TV serials, Sitcoms, Novels, Kitchen data.

Evaluation: Manually sampled accuracy over the activity frames.
Evaluation: Baselines

- No direct competitor providing activity frames.

KB Baseline: Our semantic frame (rule based) structure over the crowdsourced commonsense KB ConceptNet

Methodology Baseline: A rule based frame detector over our data and other data using an open IE system ReVerb
You open your wallet hasNextSubEvent take out money

Normalized domain: concept1 ~ verb [article] noun

Organize and canonicalized the relations as follows:

<table>
<thead>
<tr>
<th>ConceptNet 5’s relations</th>
<th>We map it to</th>
</tr>
</thead>
<tbody>
<tr>
<td>IsA, InheritsFrom</td>
<td>type</td>
</tr>
<tr>
<td>Causes, ReceivesAction, RelatedTo, CapableOf, UsedFor</td>
<td>agent</td>
</tr>
<tr>
<td>HasPrerequisite, HasFirst/LastSubevent, HasSubevent,</td>
<td>prev/next</td>
</tr>
<tr>
<td>MoratedByGoal</td>
<td></td>
</tr>
<tr>
<td>SimilarTo, Synonym</td>
<td>similarTo</td>
</tr>
<tr>
<td>AtLocation, LocationOfAction, LocatedNear</td>
<td>location</td>
</tr>
</tbody>
</table>
Methodology Baseline

Reverb, an openIE tool extracts SVO triples from text
- S and O are only surface forms.
- V is not categorized into a relation. We use a Bayesian classifier to estimate the label of V

The estimates come from MovieClips.com that provides 30K manually tagged popular movie scenes like, **action**: singing, **prop**: violin, **setting**: theater

$$P(\text{class}|\text{word}) = \frac{P(\text{class}, \text{word})}{\sum_{w_i} P(\text{class}, w_i)}$$
Methodology Baseline

Reverb, an openIE tool extracts SVO triples from text

- S and O are only surface forms.
- V is not categorized into a relation. We use a Bayesian classifier to estimate the label of V

The estimates come from MovieClips.com that provides 30K manually tagged popular movie scenes like, **action**: singing, **prop**: violin, **setting**: theater

<table>
<thead>
<tr>
<th>MovieClips tag</th>
<th>Knowlywood attributes</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>action</td>
<td>activity.v</td>
<td>cut</td>
</tr>
<tr>
<td>prop</td>
<td>activity.o</td>
<td>knife</td>
</tr>
<tr>
<td>setting</td>
<td>location</td>
<td>bar</td>
</tr>
<tr>
<td>occasion</td>
<td>time</td>
<td>thanksgiving</td>
</tr>
<tr>
<td>charactertype</td>
<td>participant</td>
<td>policeman</td>
</tr>
</tbody>
</table>
Parent | Participant | Prev | Next | Location | Time

| Knowlywood | ~1 M | High accuracy & high coverage |
| ConceptNet based | ~ 5 K | High accuracy & low coverage |
| Reverb based | ~ 0.3 M | Low accuracy & high coverage |
| Reverb clueweb | ~ 0.8 M | Low accuracy & high coverage |
Visual alignments

~30,000 Images from movies, and additionally, >1 Million images via Flickr tag matching:

- riding, road, bicycle ..
- Match verb-noun pairs from Knowlywood as *ride bicycle*
- Flickr Activity vector = road
  DOT
  Knowlywood = man, road
- ride a bicycle participant: man, boy
  location: road
External use case -1: Semantic indexing

Given: participant, location and time
Predict: the activity
Ground truth: Movieclip’s manually specified activity tag.

<table>
<thead>
<tr>
<th>System</th>
<th>MRR</th>
<th>Hit rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>ReVerbClue</td>
<td>0.070</td>
<td>0.180</td>
</tr>
<tr>
<td>ConceptNet</td>
<td>0.143</td>
<td>0.345</td>
</tr>
<tr>
<td>ReVerbMCS</td>
<td>0.254</td>
<td>0.415</td>
</tr>
<tr>
<td>Knowlywood</td>
<td>0.327</td>
<td>0.610</td>
</tr>
</tbody>
</table>

Atleast one hit in Top 10 predictions

Thank you!
Browse at bit.ly/webchild
**External use case 2: Movie Scene Search**

Mark goes hunting with Sophie’s dad. Jeremy goes hunting with Sophie’s dad. Mark tries to **kill a bird**. The man injures it simply. The man tries to break its neck.

from Sitcom: Peep Show

.. Carlos and Susan are still painting over the graffiti on the wall as those people discuss **To Kill a Mocking Bird**, however, while talking.

from TV series: Desperate Housewives

Method: A generative model encoding that a query holistically matches a scene if the participants and activity fit well with the query.

\[
P(q|s_t) = \sum_{a \in K} \sum_{p \in A_p} P(q|a) \cdot P(a|p) \cdot P(p|s_t)
\]
Conclusion

External use case -1:
Semantic indexing

Given: participant, location and time
Predict: the activity
Ground truth: Movieclip’s manually specified activity tag.

<table>
<thead>
<tr>
<th>MRR</th>
<th>Hit Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>ReVerbClue</td>
<td>0.070</td>
</tr>
<tr>
<td>ConceptNet</td>
<td>0.143</td>
</tr>
<tr>
<td>ReVerbMCS</td>
<td>0.254</td>
</tr>
<tr>
<td>Knowlywood</td>
<td>0.327</td>
</tr>
</tbody>
</table>

Atleast one hit in Top 10 predictions

Thank you! Browse at bit.ly/webchild