# Lights, Camera, Action: Knowledge Extraction from Movie Scripts

Niket Tandon
Max Planck Institute for Informatics
ntandon@mpi-inf.mpg.de

Gerard de Melo
Tsinghua University, China
demelo@tsinghua.edu.cn

Abir De
IIT Kharagpur, India
abir.de@cse.iitkgp.ernet.in

Gerhard Weikum
Max Planck Institute for Informatics
weikum@mpi-inf.mpg.de

## ABSTRACT

With the success of large knowledge graphs, research on automatically acquiring commonsense knowledge is revived. One kind of knowledge that has not received attention is that of human activities. This paper presents an information extraction pipeline for systematically distilling activity knowledge from a corpus of movie scripts. Our semantic frames capture activities together with their participating agents and their typical spatial, temporal and sequential contexts. The resulting knowledge base comprises about 250,000 activities with links to specific movie scenes where they occur.

## Categories and Subject Descriptors

I.2.7 [**Natural Language Processing**]: Text Analysis

## Keywords

Activity Knowledge; Commonsense Knowledge Acquisition

## 1. INTRODUCTION

**Motivation and State of the Art:** There is a strong need for computers having *commonsense knowledge* to support the interpretation of user input in search, dialogs, etc.. Digital assistants like Amazon Echo, Microsoft Cortana, Apple Siri or Google Now would especially benefit from knowledge about human activities. This should be in machine-readable form, e.g. as semantic frames with attributes (or slots) about participating agents and their spatio-temporal contexts. An activity such as a *romantic dinner*, for instance, takes places indoors (usually a restaurant) in the evening or at nighttime, and typically involves a romantic couple, drinks, candle light, etc., and is often succeeded by other activities like *kissing* or (alternatively) *arguing* and *breaking off with someone*.

Publicly available knowledge bases (KBs) like DBpedia, Freebase, Wikidata, and Yago and commercial KBs at Google, Microsoft, Bloomberg, etc. focus on facts about individual entities, hardly containing any commonsense knowledge at all. There are several sizable commonsense KBs, most notably, ConceptNet [7] and WebChild [9]. However, these focus on general and more "static" commonsense like concept hierarchies (subtype of, part of, member of, etc.) and properties of physical concepts (shape, color, etc.).


*WWW 2015 Companion,* May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3473-0/15/05.
http://dx.doi.org/10.1145/2740908.2742756.

Recent work in computer vision [6] has manually annotated human activities in short video clips, such as *cutting onions* in cooking scenes, with attributes such as *tool: knife*. This data has been compiled for training and benchmarking purposes and is very small and specialized. Knowledge bases about images like ObjectBank/ImageNet [5] and NEIL [2], on the other hand, focus on visual objects in (static) scenes and do not address the underlying activities.

**Goal:** Our aim is to automate the construction of semantic frames for human activities, in order to build a wide-coverage *activity KB*. This would be a valuable asset for interpreting user intentions in natural-language querying or dialogs, and also for improving the understanding of visual contents in photos and videos (e.g., as additional features for training). For example, when a user searches for outdoor kissing scenes in movies, the knowledge that these typically involve a woman, a man, and perhaps a beautiful beach, sunset, etc. can be harnessed to improve both precision and recall.

**Approach and Contribution:** Our approach to this end is to tap on movie scripts, which are available for many movies on the Internet (e.g., at `dailyscript.com`). Scripts include a clear structuring into scenes, descriptions of scene settings/locations, speakers and the full dialog, etc. Moreover, when scripts come with representative images or time points in the movie, it is possible to align a scene description with the actual visual contents. The main difficulty, however, is that the contents of movie scripts is merely in textual form — still very far from a structured KB representation.

Our pipeline for information extraction is based on semantic parsing methods (see [1] for an overview). A major task then is to map the slot values of these frames (activity type, location, participating agent, etc.) onto proper disambiguated word senses, which we address using strong priors fed into an integer linear program. A second major building block of our method is the inference of predecessor and successor activities. For this, we have devised an algorithm based on frequent sequence mining.

We applied this methodology to an input corpus with 560 movie scripts with a total of 148,296 scenes. The constructed activity KB comprises 244,789 different activities, each represented by a frame that identifies the participating agents, the place and time of the activity, and preceding and succeeding activities, such as *romantic dinner* followed by *kissing*, *wedding*, etc. Most of these activity frames are also linked with video scenes where the activities occur. The activity KB can be browsed at `tinyurl.com/activitykb`.

## 2. METHODS

**Semantic Parsing:** We devised a customized semantic parsing pipeline that starts with the raw input scripts, performs information extraction, disambiguates constituents (the potential attribute values of an activity), all the way to constructing a frame structure for candidate activities. We process the input data scene by scene, where we use simple cues for splitting a script into scenes. Each

**Table 1: Anecdotal example results.**

| Activity | Parent | SimilarTo | Participant | Prev | Next | Location | Time |
|---|---|---|---|---|---|---|---|
| open#1<br>door#1 | open#1<br>barrier#1 | shut#1<br>door#1 | person#1 | knock#2<br>door#1 | tell#2<br>man#1 | room#1 | day#4 |
| threaten#2<br>a woman#1 | warn#1<br>a person#1 | warn#1<br>a nurse#1 | man#1 | take#16<br>picture#1 | end#2<br>relationship#1 | home#2 | night#1 |

input sentence is tokenized, POS-tagged and chunked. The sentence is split into sub-phrases using the clause structure of the sentence with the ClausIE tool [3]. We then run OpenNLP (`opennlp .sourceforge.net`) for chunking each phrase.

**Sense Disambiguation:** In order to distinguish different senses of words, we use the IMS tool [10] to map words onto their WordNet senses. Virtually all such tools operate at word granularity, though, and do not handle multi-word phrases. To overcome this limitation, we identify and disambiguate the head word in each noun phrase. For example, we map `the moving bus` to "*bus#1*", where "*bus#1*" refers to the first sense of the word `bus`) in WordNet: the vehicle sense. We apply the same heuristics to verbal phrases, mapping, for example, `begin to shoot` in the sentence "he began to shoot a video in the moving bus" onto "*shoot#2*" that is, killing someone. This is obviously wrong (the correct sense would be "*shoot#4*": filming). We correct such mistakes by jointly disambiguating verb phrases and the noun phrases for their arguments, using a judiciously designed integer linear program (ILP). We use the state-of-the-art ILP solver Gurobi (`www.gurobi.com`) for computing the solution. Details are omitted for lack of space.

**Inferring Attributes:** The previous step already yields a preliminary but noisy frame structure. We employ additional inference steps for further cleaning and eliminating overly noisy outputs.

As activities are primarily expressed by verbal phrases, we link the WordNet verb sense of the previous step with VerbNet [4], a manually curated high-quality linguistic resource for English verbs, which is already aligned with WordNet. VerbNet provides syntactic information (e.g., the number of objects that a verb can or should have: 0, 1, or 2) and argument restrictions for verb senses. For example, for the verb sense `shoot#2` (killing), the role restriction is `Agent.animate V Patient.`**`animate`** `PP Instrument.solid` where `animate` refers to living beings, as opposed to inanimate objects. With the joint mapping of verbs and their arguments onto senses, we can infer that this `shoot#2` sense is not compatible with the argument "the video", as it is not animate. This way, we can disqualify the incorrect interpretation of "shoot". We only accept candidate frames that satisfy these kinds of semantic argument restrictions. An example output of our pipeline for semantic parsing and frame construction is shown in Table 2.

**Table 2: Semantic parse: "*he began to shoot a video in the moving bus*"**

| Phrase | WordNet Mapping | VerbNet Mapping | Expected Frame |
|---|---|---|---|
| the man | man#1 | Agent . animate | Agent: man#1 |
| begin to shoot | shoot#4 | shoot#vn#3 | Action: shoot#4 |
| a video | video#1 | Patient . solid | Patient:video#1 |
| in | in | PP . in | |
| the moving bus | bus#1 | NP . Location . solid | Location:moving bus#1 |

**Inferring Activity Order:** Given the noisy sequences of activities in scenes obtained so far, we distill these by running an algorithm for generalized sequence pattern mining based on [8]. An activity $a_1$ follows $a_2$ with a score proportional to the support $\frac{\text{freq}(a_1 \text{ directly follows } a_2)}{\text{freq}(a_1)\ \text{freq}(a_2)}$. We accept a precede/succeed relation between two activities if this score is above a specified threshold.

**Linking to Visual Scenes:** We attach key frames in videos to activities. For this task, we harness subtitles in the video footage. and match these against characteristic text phrases in the dialog of a movie script. If available, we also use timestamps for fine-tuning this alignment between script and video.

# 3. RESULTS

From the input of 560 movie scripts with a total of 148,296 scenes, we have constructed an activity KB with 244,789 activity frames. These are organized into a subsumption hierarchy, and each frame has attributes like participating agents, typical location, typical daytime, predecessor frame, successor frame — sometimes only partially filled. The only prior KB that had some knowledge of this kind is ConceptNet 5 [7]. However, it has only 28,273 concepts of this kind, with about 59,168 instances of precede/succeed attributes. Moreover, all these entries are at the surface-word level, none are disambiguated, and there is no linkage to visual contents.

For a preliminary evaluation of the quality of the distilled knowledge, we sampled the data in our activity KB along three dimensions: i) Are the activity type itself and the participating agents appropriate and are their mappings to WordNet senses correct? ii) Are the preceding and succeeding activity types appropriate? iii) Are the activity frames linked to scenes where the activity actually occurs? We manually evaluated 100 samples for each of these evaluation tasks. We found that the precision (i.e., fraction of correct output) is reasonably high: 84% ($0.84 \pm 0.02$) for the first question, 83% ($0.83 \pm 0.08$) for the second, and 78% ($0.78 \pm 0.06$) for the third. For statistical significance, we computed Wilson score intervals for $\alpha = 95\%$. Assessing the recall of the activity KB requires a more sophisticated setup and is subject of ongoing work. We are devising additional cleaning procedures to further improve the precision, and exploring extrinsic use-cases of the KB.

Table 1 illustrates a few anecdotal samples. The complete activity KB is accessible at `tinyurl.com/activitykb`.

# 4. REFERENCES


[1] Y.Artzi, N. FitzGerald, L.S. Zettlemoyer: Semantic Parsing with Combinatory Categorial Grammars. ACL 2013

[2] X. Chen, A. Shrivastava, A. Gupta: NEIL: Extracting Visual Knowledge from Web Data. ICCV 2013

[3] L. Del Corro, R. Gemulla: ClausIE: Clause-based Open Information Extraction. WWW 2013

[4] K. Kipper, A. Korhonen, N. Ryant, M. Palmer: A large-scale classification of English verbs. LREC 2008

[5] L.-J. Li, H. Su, Y. Lim, F.-F. Li: Object Bank: An Object-Level Image Representation for High-Level Visual Recognition. Int. J. of Computer Vision 107(1), 2014

[6] M. Rohrbach, S. Amin, M. Andriluka, B. Schiele: A Database for Fine Grained Activity Detection of Cooking Activities. CVPR 2012

[7] R. Speer, C. Havasi: Representing General Relational Knowledge in ConceptNet 5. LREC 2012

[8] R. Srikant, R. Agrawal: Mining Sequential Patterns: Generalizations and Performance Improvements. EDBT 1996

[9] N. Tandon, G. de Melo, F.M. Suchanek, G. Weikum: WebChild: Harvesting and Organizing Commonsense Knowledge from the Web. WSDM 2014

[10] Z. Zhong, H.T. Ng: It Makes Sense: A Wide-Coverage Word Sense Disambiguation System for Free Text. ACL 2010