Learning the Central Events and Participants in Unlabeled Text

Nathanael Chambers NCHAMBER@USNA.EDU
United States Naval Academy, Annapolis, MD 21402 USA

Dan Jurafsky JURAFSKY@STANFORD.EDU
Stanford University, Stanford, CA 94305

Abstract

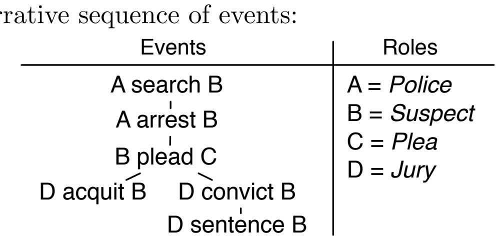
The majority of information on the Internet is expressed in written text. Understanding and extracting this information is crucial to building intelligent systems that can organize this knowledge. Today, most algorithms focus on learning atomic facts and relations. For instance, we can reliably extract facts like "Annapolis is a City" by observing redundant word patterns across a corpus. However, these facts do not capture richer knowledge like the way detonating a bomb is related to destroying a building, or that the perpetrator who was convicted must have been arrested. A structured model of these events and entities is needed for a deeper understanding of language. This talk describes unsupervised approaches to learning such rich knowledge.

1. Overview

This talk describes a new approach to knowledge acquisition and extraction that learns rich structures of events (e.g., plant, detonate, destroy) and participants (e.g., suspect, target, victim) over a large corpus of news articles, beginning from scratch and without human involvement. Early models in Natural Language Processing (NLP) relied on similar high-level representations like frames and scripts (structured representations of events, their causal relationships, and their participants) to drive interpretation of syntax and word use. Scripts, in particular, were central to research in the 1970s and 1980s (Schank & Abelson, 1977). However, scripts were hand-coded, unable to generalize to new domains. Modern statistical approaches and advances in NLP now enable new representations and large-scale learning over many domains.

The main problem with scripts and similar common-sense knowledge structures was that the need for hand construction, specificity, and domain dependence prevented robust language understanding. The many diverse and varied situations in the world cannot realistically be hand coded for every language application and domain. For example, a script about "eating in a restaurant" (the most famous Schankian script) cannot assist the understanding of situations like corporate acquisitions and football games. The development of scripts proved too time intensive and too brittle when changing contexts.

In this talk, I describe attempts to learn script-like information about the world, including both event structures and the roles of their participants, but without pre-defined frames, roles, or tagged corpora. Consider the following *event schema*, informally represented. The events on the left follow a set of participants through a series of connected events that constitute a narrative sequence of events:



Being able to robustly learn sets of related events (left) and frame-specific role information about the argument types that fill them (right) could assist a variety of NLP applications, including summarization, question answering and information extraction (IE). I will briefly describe its successful application to IE below.

Extracting structured knowledge about the world from raw text is a relatively new goal in NLP. The majority of machine learning approaches have focused on shallower representations of meaning. Fact and relation extraction has focused on extracting true state-

Appearing in *Proceedings of the 29^{th} International Conference on Machine Learning*, Edinburgh, Scotland, UK, 2012.

ments about the world, typically in the form of relation triples: *A relation B*. Examples from the Open IE system (Banko, 2009) include, *Napoleon married Josephine*, *Einstein born in Ulm*, and *oranges contain Vitamin C*. The system learns synonymous patterns (e.g., sign on, enter, and join), as well as classes of similar concepts (e.g., company, business, inc., organization) by observing patterns with the same argument types. Other approaches vary in their use of (or lack of) seed examples, domain text, and target knowledge (Kok & Domingos, 2008; Carlson et al., 2010; Huang & Riloff, 2010). These are largely unsupervised.

The event schemas in this talk are a complimentary, but different type of knowledge from the above relations/facts. Event schemas represent sets of *dependent* relations and entities in specific roles, while the above systems learn lists of *independent* relations. This work is thus unique in learning relations across relations, or imposing a structure over otherwise independent relations. Further, my algorithm does not capture specific facts, like napoleon married josephine, but instead captures general events/relations like People marry People. It then learns the related events and roles such as meet, date, bear children, retire, etc.

Briefly, the key to learning rich event knowledge is identifying syntactic clues that connect events. My learning algorithm focuses on entities in text (e.g., people, places, things) and uses repeated mentions of the same entity to link related events. For instance, the phrases 'Bruce pushed the man' and 'the man fell down' convey implicit knowledge about the world, namely, that *push* and *fall down* are related. We know this because *the man* appears with both verbs. We can thus loosely conclude that the two events occur together in the real world. I identify these connections between events over a million document corpus and learn event structures for thousands of verbs. The result is a connected graph with events as nodes, and mutual information scores as edges.

Given this graph, we cluster events using the edge weights as defined by the entities we observed. Figure 1 shows three possible clusters extracted from the graph. Each cluster represents a unique entity, defined over the same events, and can be merged into one event schema representation.

The algorithm learns schemas for over 1,800 English verbs over a range of domains. However, the end goal is not to simply learn knowledge for knowledge's sake. I have successfully applied event schemas to an IE task that traditionally required *supervised* learning. Briefly, the task is to extract entities that performed a known action (e.g., bombings). We are the first to

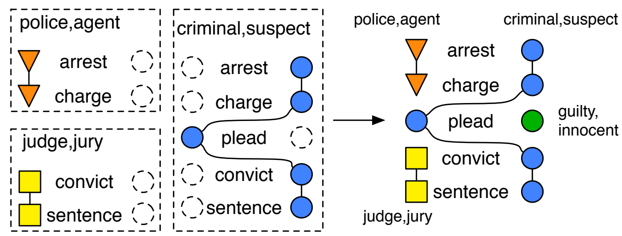


Figure 1. Three separate event clusters, learned by following entities, mapped to one coherent event schema.

attempt this task without knowing the actions in advance, and without labeled data. The system had to first learn that bombings and kidnappings exist in the world from raw text. It then learned that each is characterized by a series of events and unique entities like perpetrators, targets, and weapons. Below is one such *event template* that was learned automatically.

Bombing Template
{detonate, blow up, plant, explode, defuse, destroy}

Perpetrator:	Person who detonates, plants, blows up
Instrument:	Object that is planted, detonated, defused
Target:	Object that is destroyed, is blown up

All previous work on this task depends on knowing ahead of time that perpetrators exist, as well as human annotations of example entities. I instead induced this event knowledge, and then extracted entities as is standard in the task, achieving accuracies approaching those of supervised learning algorithms.

Acknowledgments

Thanks to the original support of the NLP Group at Stanford University, and the fresh support of Nathanael's new institution, the United States Naval Academy.

References

Banko, Michele. *Open Information Extraction for the Web*. PhD thesis, University of Washington, 2009.

Carlson, Andrew, Betteridge, J., Kisiel, B., Settles, B., Jr., E.R. Hruschka, and Mitchell, T.M. Toward an architecture for never-ending language learning. In *Proceedings of AAAI*, 2010.

Huang, Ruihong and Riloff, Ellen. Inducing domain-specific semantic class taggers from (almost) nothing. In *Proceedings of the Association for Computational Linguistics (ACL)*, 2010.

Kok, Stanley and Domingos, Pedro. Extracting semantic networks from text via relational clustering. In *Proceedings of ECML*, 2008.

Schank, Roger C. and Abelson, Robert P. *Scripts, plans, goals and understanding*. Lawrence Erlbaum, 1977.