Research Statement

Niranjan Balasubramanian

November 27, 2013

My research spans two broad areas: Information retrieval and Natural Language Processing. Applications like Question answering, event extraction and summarization systems, and web search providing easy access and organization to the vast amounts information on the web. These challenging applications require computer systems to extract, understand and reason with knowledge present in natural language texts. This long-standing Artificial Intelligence (AI) vision has tremendous societal and scientific impacts. My research is motivated by this vision and aims at building large scale open domain knowledge targeted towards specific applications.

Consider a system answering a 4th grade science question: "Which is the best conductor of electricity? (A) metal fork (B) rubber boat". The system needs access to the knowledge that i) a metal fork is made of metal, ii) metals conduct electricity, and iii) properties of a metal apply to things made of metal and reason with this knowledge to arrive at the answer. Consider an event extraction system that is "reading" an article about an arrest event. If the system had knowledge about what a typical arrest event is – i.e., who the key actors are and what their roles are – then it can use that information to identify the salient pieces to extract from the article.

My current research focuses on developing methods for extracting such large scale open-domain knowledge and robust mechanisms for applying this knowledge. For wide applicability, I target methods that meet the following design goals: Scale to arbitrary domains, model semantics without ambiguity, and generalize to unseen contexts. I use Open Information Extraction (Open IE) techniques to extract open-domain relations and representations augmented with semantic classes to reduce ambiguity and improve generalization.

My research methods are centered on identifying the core aspects of the problem space through careful analysis of existing systems and data. For instance, in my work on event schemas, I analyzed output from previous system and identified the key shortcomings of underspecified representations. The insights led to a better representation that vastly improved the quality of the schemas. In my work on mobile search, I conducted a systematic study of data transfer costs in cellular networks. This study identified a key energy inefficiency, which spawned a large body of research aimed at minimizing its impact. [1]

Here I describe some of my research and lay out my vision for future work.

NLP [EMNLP 2013, AKBC-WEKEX 2012, 2013]

Modeling Events

Event schemas that specify actors and their roles within events are widely used in event extraction. Figure 1 shows an example arrest schema. The key actors are an arresting agent who arrests and charges a suspect, a lawyer who represents the suspect and a judge who rules on the case. My research aims to automatically generate these schemas from text with no manual effort with a specific focus on generating coherent schemas.

The main premise behind my work is that an accurate model of co-occurring actors and their actions can provide a basis for automatically generating schemas. The main challenge is in defining a suitable representation for the actions. My analysis of the output from a previous system showed that simpler (subject, verb) and (verb, object) pairs are underspecified representations that split

[1] This work has more than 350 citations.

critical context. In response, I developed an Open IE triple-based solution that uses a (Arg1, Relation, Arg2) triple that captures more specific information about the actions and reduces ambiguity. However, this reduction in ambiguity comes at a cost of increased sparsity and reduced generalization. To counter this issue, I represent arguments using semantic classes, which allows the model to generalize beyond the specific entities to new unseen contexts. The relational co-occurrence model (Rel-grams) built over this triple representation yields a form of entailment type knowledge [7] and produces schemas that are more coherent than state-of-the-art systems [8].

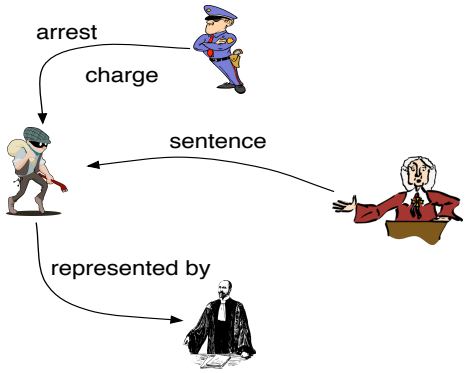

Figure 1: Arrest schema: A model for an arrest scenario including key actors, the police, the suspect, judge etc. and their roles.

Question Answering

Achieving human-level performance on tasks that require intelligence has a long tradition in the history of AI. As one of the foundational members of the Allen Institute for Artificial Intelligence, I am co-leading efforts to design and develop a QA system that is capable of passing a 4th grade science exam. [2] This is an exciting long-term research project that seeks to address several fundamental challenges in representation, extraction and reasoning all in the context of a single task.

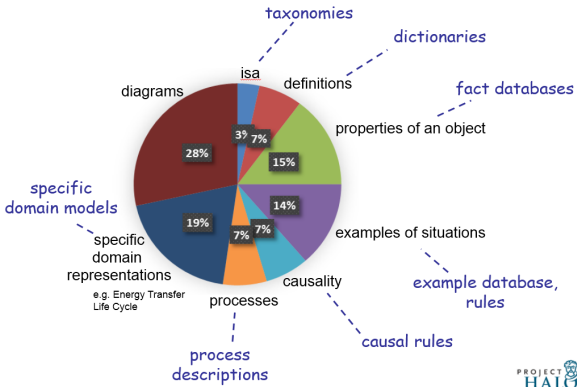

Figure 2: Knowledge Requirements for passing a 4th Grade Science Exam

As a first step in this task, we studied the knowledge requirements for passing a 4th grade science exam [9]. The knowledge requirements summarized in Figure 2 shows that this is a challenging task requiring a wide range of knowledge and robust reasoning methods. In addition to factual knowledge (e.g., iron conducts electricity), we identify three other types of useful knowledge: 1) Definitional knowledge, 2) Domain knowledge expressed via general purpose relations such as cause/effect, entity/function, 3) Implications representing domain and background knowledge (e.g., animal breathes oxygen $-$enables$\rightarrow$ animal make energy), and 4) Qualitative domain models (e.g., Reasoning with predator-prey models: If population of snakes rise, what happens to the population of frogs?).

We are building a wide array of solutions to address these difficult challenges. We extract definitional and general purpose relations such as cause/effect and entity/function using hand-generated lexico-syntactic patterns that exploit strong regularities in language. We use Open IE style relations to represent information but expand them to cover nested relations. In addition, we also use facts extracted using a state-of-the-art Open IE extractor. Our preliminary experiments show about 15% improvement over simpler BOW baselines.

Information Retrieval [SIGIR 2010, CIKM 2009, IMC 2009, CoNext 2012]

Combining Alternatives

The IR research community continuously develops query representations, retrieval models, and various ranking algorithms. As part of my thesis, I developed a dynamic query-dependent approach for combining different alternatives. The main premise behind my thesis is that different alternatives work well for different queries. For example, a navigation query (intent to visit a specific url) is well served by user click based features, whereas a informational query is better served by query-document

[2] I also contribute to the long term research planning efforts and earlier co-wrote grant proposals with Dr. Oren Etzioni to obtain funding on earlier versions of this project.

match features. If we can select the best choice(s) for a given query, then we can further improve retrieval performance.

To this end, I developed a novel method that estimates the relative performance of the alternatives with respect to a baseline using easy to compute retrieval features [6]. The key insight behind this method was that accurate estimation of the absolute effectiveness value was not essential. The estimation only needed to induce a good ordering of the available alternatives. This relative estimation method outperformed using the single best alternative for query representations [4] and standard fusion methods for combining ranking functions [5].

Topic Pages

As an alternative to the typical web search paradigm, I built a system that automatically generates wikipedia like pages for queries [3]. The primary challenge here is to identify the salient aspects pertaining to the query topic. I used web search logs to build diverse aspect models on topics. To generalize to topics beyond those that are observed in the search logs, I generalized the aspect models to include information from related topics. A second challenge here is to extract and organize information pertaining to the diverse aspects in a coherent fashion. I built a sentence extractor that identifies most typical connection between the topic and its aspect and used simple word-precedence models to organize the retrieved sentences. The resulting topic pages outperformed state-of-the-art summarization systems in terms of grammaticality, salience and coherence.

Mobile Search

I studied the impact of system constraints on web search from mobile phones. I conducted a systematic study of how network activity consumed energy in mobile phones. My study revealed that energy consumption also depended on inter-transfer times in addition to the size of the data being transferred [2]. Based on this insight, I designed interaction strategies that reduced the energy consumption of web search applications. In addition to improving the energy efficiency of web search, the key finding in this work led to a large body of work aimed at addressing the energy inefficiency.

I also worked on *FindAll*, a mobile search engine aimed at improving local availability of previously visited documents. Because indexing on the phone is expensive, the system must balance local availability against resource usage and energy consumption. Using actual search logs from mobile users, we learned user-specific re-finding patterns to predict when a user is likely to re-find documents. Using this predictive model, FindAll selectively indexes documents when cost of indexing is lower than cost of re-finding the document over the network. Evaluations show that FindAll dramatically improves local availability for heavy users without increasing the energy costs.

Future Work

I am interested in extracting open-domain knowledge from text. In particular I want to extracting script-like descriptions of open-domain events and processes. My past work on Rel-grams and open event schemas provides a starting point but much work remains do be done. I am also interested in methods that can handle the gaps and noise inherent in such automatically extracted knowledge.

Script-like Knowledge

Scripts are general purpose descriptions of events or scenarios. They include the key actors, their actions, and causal and temporal relationship between the different actions [11]. A dinner script for example is an ordered sequence of actions: An actor going to a restaurant, placing an order with a waiter, eating the food, and then paying the bill.

If we read the sentences "John went to Bill's restaurant and ordered steak. [He] paid $50 for the meal.", we easily infer that [He] refers to John and that John most likely ate the steak. Scripts provide valuable general purpose knowledge that can fill in these missing connections but aggregating high-quality scripts is a challenging task because many aspects of this knowledge are often implicit and harder to detect. I am interested in methods that can aggregate and generalize from explicitly mentioned aspects.

For instance, causal links can be explicit (e.g., John ate lunch early *because* he was hungry) or implicit (e.g., John ate lunch early. He was hungry). The explicit mentions can be identified more reliably using lexical causal connectives (e.g, *because*) but the implicit links are harder. There are other structural constraints such as transitivity (i.e., A causes B and B causes C $\rightarrow$ A causes C) which can further improve precision. The bootstrapping task here is to detect explicit links with high-precision and then learn patterns over generalized representations of the actors and their actions.

Scripts also serve as a template for extracting salient information about events by specifying the key actors and their actions. My past work on open event schemas produced coherent models of events in terms of the key actors and their roles and thus provide a strong basis constructing high-quality event extractors at scale. Expanding schemas to include extractors can also be viewed as a bootstrapping procedure starting from a strong high-quality model of an event. As with any bootstrapping approach the challenges are in finding appropriate sources for expansion and avoiding drifting from the source model.

Reasoning with Automatically Extracted Knowledge

Reasoning with knowledge extracted from texts require robust mechanisms to handle the inherent uncertainty, redundancy, and vocabulary mismatch issues. Previous textual entailment approaches that simply combine a large number of weak features. In contrast, I wish to explore approaches that preserve the deductive style of inference as much as possible, but fall back to weaker methods when necessary.

To this end, I am interested in building a probabilistic reasoning framework. Specifically, I want to model question answering as a marginal inference problem in a Markov Logic Network (MLN) with text-based implications as first-order rules and Open IE style relations as evidence. Different from traditional applications of MLNs, the knowledge and rules here are textual, which result in gaps from vocabulary mismatch. To address these gaps, I will build a search solution that uses coarse but fast methods to locate plausible gaps and bridge them using deeper (more expensive) textual entailment techniques. The key challenges here include keeping the search space tractable and to have broad coverage entailment methods to handle lexical variations.

Information Retrieval

I am also interested in exploiting semantic knowledge for Information retrieval. In the past, IR applications have had mixed success with using semantic resources. The key limiting factor was the coverage of the resources used and scalability of the methods. Recent advances in large scale language processing and knowledge extraction techniques provide an ideal opportunity to test integration of semantics. Open Information Extraction presents an ideal starting point. It provides fast and a shallow representation of salient information in a corpus that can be used to improve retrieval.

References

[1] Aruna Balasubramanian, Niranjan Balasubramanian, Samuel J Huston, Donald Metzler, and David J Wetherall. Findall: a local search engine for mobile phones. In *Proceedings of the 8th international conference on Emerging networking experiments and technologies*, pages 277–288. ACM, 2012.

[2] Niranjan Balasubramanian, Aruna Balasubramanian, and Arun Venkataramani. Energy consumption in mobile phones: a measurement study and implications for network applications. In *Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference*, pages 280–293. ACM, 2009.

[3] Niranjan Balasubramanian and Silviu Cucerzan. Topic pages: An alternative to the ten blue links. In *Semantic Computing (ICSC), 2010 IEEE Fourth International Conference on*, pages 353–360. IEEE, 2010.

[4] Niranjan Balasubramanian, Giridhar Kumaran, and Vitor R Carvalho. Exploring reductions for long web queries. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 571–578. ACM, 2010.

[5] Niranjan Balasubramanian, Giridhar Kumaran, and Vitor R Carvalho. Learning to select rankers. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 785–786. ACM, 2010.

[6] Niranjan Balasubramanian, Giridhar Kumaran, and Vitor R Carvalho. Predicting query performance on the web. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 785–786. ACM, 2010.

[7] Niranjan Balasubramanian, Stephen Soderland, Mausam, and Oren Etzioni. Rel-grams: a probabilistic model of relations in text. In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction*, pages 101–105. Association for Computational Linguistics, 2012.

[8] Niranjan Balasubramanian, Stephen Soderland, and Oren Etzioni Mausam. Generating coherent event schemas at scale. In *Proceedings of the Empirical Methods in Natural Language Processing*. ACM, 2013.

[9] Peter Clark, Phil Harrison, and Niranjan Balasubramanian. A study of the akbc requirements for passing an elementary science test.

[10] S. Patwardhan and E. Riloff. A unified model of phrasal and sentential evidence for information extraction. In *Proceedings of EMNLP 2009*, 2009.

[11] Roger C Schank and Robert P Abelson. *Scripts, plans, and knowledge*. Yale University, 1975.