Research Statement

Niranjan Balasubramanian

December 20, 2013

My research spans two broad areas: Information retrieval and Natural Language Processing. Applications such as web search, information extraction, summarization and question answering provide easy access 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. It also needs to 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.

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 specify actors and their roles within events. They 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 without any manual effort.

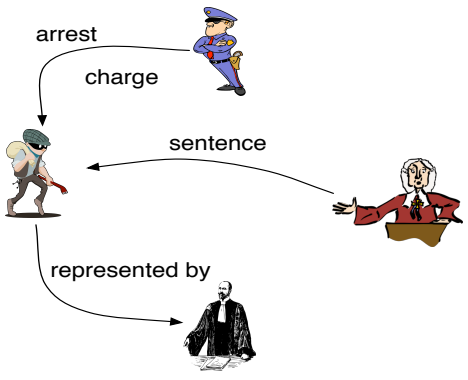

Figure 1: Arrest schema: A model for an arrest scenario including key actors, the police, the suspect, judge etc. and their roles.

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 and split 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].

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. [1] 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.

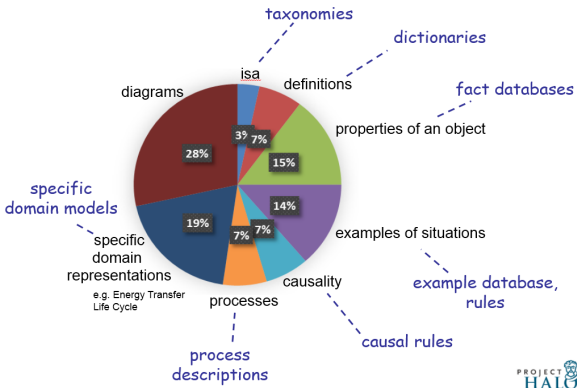

Figure 2: Knowledge Requirements for a 4th Grade Science Exam

As a first step, 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 simple keyword search 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 click-based features, whereas a informational query is better served by query-document 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 [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].

[1] I also contribute to the long term research planning and have written grant proposals with Dr. Oren Etzioni that secured funding on earlier versions of this project.

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 but diverse 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

Building systems that can understand and reason with natural language texts is one of the central visions of AI. I am interested in three veins of research that build towards this vision:

1. Acquiring knowledge – Extracting and understanding information in texts itself requires background knowledge. For example models of salient aspects of certain types of entities (e.g., actors *act in* movies, actors *win* awards) can help extract and organize information about people. Similarly, rich descriptions of events or scenarios, processes and their interactions can help in extracting and understanding information about events from texts. Scalable methods for acquiring such background knowledge is essential for extracting richer structures from text.

2. Reasoning with knowledge from texts – Much of world's knowledge is already available in the form of natural language texts. While extraction capabilities are improving, they have several issues that need to be addressed in order to be useful for end applications.

3. Semantic search capabilities – Advent of scalable extraction and language processing capabilities present a great opportunity to push information retrieval toward more semantic representations and retrieval models.

Below I describe three specific problems that will be part of my initial efforts in these areas.

Script-like Knowledge for Modeling Events

Understanding natural language text requires broad coverage open-domain knowledge. For example, "John went to Bill's restaurant. [He] ordered a steak. [He] paid $50 in cash for [the meal]". When we read these sentences, we easily identify that the key actors are John, the restaurant, a waiter who took the order etc. Note that the waiter was never mentioned in the text. Also, we are able to infer that the pronouns in the sentences all refer to John, [the meal] refers to the steak, and going a step further we can even infer that John probably ate the steak before paying for the meal. To do this automatically, we need a rich model of script-like knowledge about people eating in restaurants.

Script-like knowledge serve as general purpose description of events or scenarios. They include the key actors, their actions, and causal and temporal relationship between the different actions. My prior work on open event schemas provides a starting point for building scripts but many challenges remain. Schemas provide a high precision model of scenarios but do not have associated extractors – i.e., patterns that can be used to extract from new texts. Scripts also need to resolve entity and event co-references. Third, scripts must include causal and temporal ordering of the actions in a scenario.

I view these challenges as a general problem of extrapolating and generalizing from knowledge that can identified with high precision. Development of high precision methods as well as representations that allow for generalization of the knowledge to new unseen contexts. Building extractors can be viewed as an expansion or generalization of the actions and their actions. Large scale semantic resources such as Freebase, YAGO and NELL can be exploited for this expansion. Resolving co-reference, and identifying causal and temporal links are challenging because the links can be implicit. However, the explicit links can be identified more reliably especially when combined with structural constraints such as transitivity and can be aggregated to identify implicit links.

Reasoning with Automatically Extracted Knowledge

Reasoning is critical to the AI vision and is necessary for several challenging tasks like question answering. As mentioned earlier, question answering, even at a 4th grade level, requires reasoning over a wide variety of knowledge. The required knowledge ranges from facts expressed as simple relations (e.g., An iron nail *is composed of* iron) to more complex inference rules (e.g., x *is* metal $\rightarrow$ x *conducts* electricity), most of which can be extracted automatically from text.

Reasoning with knowledge constructed from text presents many challenges including uncertainty, incompleteness, redundancy, and vocabulary mismatch. Unlike pure logic-based approaches, probabilistic frameworks such as Markov Logic Networks (MLN) can handle uncertainty in knowledge. However, purely deductive MLN formulations also fail because of gaps in knowledge and vocabulary mismatch issues. The popular alternative in textual entailment is to use purely feature-based entailment methods, which loose the explanatory powers of the deductive formulations.

I will explore hybrid approaches that preserve the deductive reasoning as much as possible but revert to feature-based entailment reasoning when necessary. For instance, I am currently working on an reasoning framework that uses textual inference rules and evidence in an MLN formulation but proactively detects and bridge plausible gaps using feature-based textual entailment techniques. This combines the best of both worlds – explaining the reasoning, showing where the current knowledge is lacking or where textual reasoning is required, while also retaining the robustness of the entailment style methods. The key challenges here include keeping the search space tractable and to have broad coverage entailment methods.

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.

Conclusion

I am fascinated by natural language and our own robust ability to use it to receive, communicate, and organize our ideas. I am motivated by the challenges of building systems with similar abilities. I aim to develop a research program that tackles these challenges at scale in the context of well defined target applications.

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.