RESEARCH STATEMENT

Yu Su (ysu@cs.ucsb.edu)

Data science has the potential to reshape many sectors of the modern society. This potential can be realized to its maximum only when data science becomes democratized, instead of centralized in a small group of expert data scientists. However, with data becoming more massive and heterogeneous, standing in stark contrast to the spreading demand is the growing gap between end users and data: Every type of data requires extensive specialized training, either to learn a specific query language or a data analytics software. Towards the *democratization of data science*, the overarching goal of my research is to bridge the gap between users and data, and make it easier for users who are less technically proficient to access the data analytics power needed for on-demand decision making.

The data analytics stack consists of layers of capabilities ranging from low-level data processing, data querying, to high-level data analytics like visualization and prediction, many of which require specialized expertise and are therefore largely inaccessible to non-expert users. In the broad area of data mining, natural language processing, and machine learning (deep learning in particular), the central theme of my research is to develop intelligent systems to bridge non-expert users to these powerful capabilities. Firstly, I develop *natural language interface for data querying*, that allows users to query massive and heterogeneous data via a unified natural language interface (§1). For example, users can directly ask questions like "*what's the best-selling smartphone in 2015*" or "*list popular treatments for depression to female in their twenties*," and get precise answers from the underlying data, be it knowledge bases, relational databases, electronic medical records, etc. Secondly, a large amount of data comes in unstructured textual form, and it is a grand challenge of data science to analyze massive text data. I develop *knowledge harvesting* techniques to mine structured, actionable knowledge from massive text data, which can greatly facilitate downstream analytics (§2). For example, by extracting entities like drugs and diseases and their treatment relationship from biomedical literature, we may support information needs like "*find medications to treat depression developed in the US*."

Moving forward, I plan to continue my research and develop *virtual assistants for data analytics*, that serve as a unified natural language interface to query, manipulate, and analyze data of heterogeneous types from different sources (§3). With such virtual assistants, users can stay focused on high-level thinking and decision making, instead of overwhelmed by low-level implementation details — "*Let machines understand human thinking. Don't let humans think like machines*." As a nascent filed, data science can evolve along many different paths. As a researcher and educator in data science, I believe the best path is the one that allows people of diverse gender, ethnic group, and social class to all enjoy data analytics capabilities and benefit from data.

1 Natural Language Interface for Data Querying

On-demand decision making requires the agility to interact with data (e.g., knowledge bases and relational databases) in a process of dynamically generating questions (hypotheses), as and when needed, and quickly getting answers. However, non-expert users can hardly enjoy such agility. Learning formal query languages like SQL (Structured Query Language) takes a large amount of time. On the other hand, whereas canned query forms (e.g., pre-defined query templates on graphical user interfaces) shield users from programming, they lack the necessary expressive flexibility for ad-hoc data querying. The gap between user and data is growing further with the rapidly increasing data volume and heterogeneity. For example, modern knowledge bases like Google Knowledge Graph contains thousands of entity and relation types, millions of entities, and billions of relational facts about entities; writing a formal query for such complex data is a challenge even for programming-proficient users, let alone users who are novice in programming. It may also not be economical to write ad-hoc, one-off formal queries to meet every dynamically generated information need. *Is it possible to build intelligent systems to automatically transduce user information needs into the corresponding formal queries?*

I develop natural language interfaces (NLIs) to transduce natural language commands from human users into

computer-understandable formal languages. With successful applications on a wide range of data types including billion-scale knowledge bases, million-scale semi-structured tables, and web APIs (application programming interfaces), my research advances natural language interface from *interdisciplinary* perspectives. From the perspective of natural language processing, I study both theoretical and practical aspects of natural language understanding. From the perspective of databases, I construct large-scale, high-quality benchmark datasets to evaluate state-of-the-art systems. From the perspective of human-computer interaction, I investigate different interaction mechanisms between users and data with human subject experiments.

1.1 Natural Language Interface to Knowledge Bases

Knowledge base has been a core store of knowledge for intelligent agents since the very early age of artificial intelligence (AI). Benefiting from the advances in areas like semantic web and information extraction, recent years have seen the emergence of many large-scale knowledge bases, both for general purposes such as the Google Knowledge Graph and DBpedia, or for specialized purposes such as Robo Brain (a multimodal knowledge base for robots). The capability of interpreting natural language commands over knowledge bases can benefit both humans and intelligent agents like robots. For example, backed by the Google Knowledge Graph, users can now get direct and precise answers in Google search to questions like "*who are the perpetrators of 9/11*," instead of reading lengthy webpages to dig out the answers by themselves.

Natural language interface to knowledge bases, also known as knowledge based question answering or semantic parsing, has drawn wide interests in both academia and industry, and a large number of systems have been developed. Benchmarks are indispensable for comprehensively evaluating and fairly comparing different systems. We have seen many successful examples of how robust and meaningful benchmarks can greatly expedite the development of a research area, such as the TPC benchmark suit for relational databases. We developed the first *multidimensional* benchmark for NLI to knowledge bases [2]. The benchmark consists of questions with a varying degree of difficulty along several dimensions, such as structural complexity (the number of relations involved), functions (quantitative analysis like counting or superlatives), and paraphrasing (asking the same question in syntactically-divergent ways). Examining systems under these different lens gives a comprehensive picture of the capabilities and incapabilities of each system, and provides clear insights for developing better systems. For example, *in light of our analysis on the incapability of current systems on paraphrasing, researchers from University of Edinburgh and Stanford developed a mechanism to specifically improve the robustness of NLIs to paraphrasing* [14]. This work also attracted interests from U.S. Army Research Lab (ARL), who is now funding and collaborating with us to construct the next version of the benchmark, targeting 20 times larger in size and more evaluation dimensions.

I developed an array of models, including answer type prediction [6], user feedback [1], question revision [7], answer triggering [8], and transfer learning [3], to make natural language interface more accurate, robust, and usable. For example, we conducted the first study on transfer learning in order to build NLIs that can simultaneously handle multiple knowledge bases of different domains [3]. Transfer learning in natural language interface is a particularly intriguing but difficult problem, because different domains involve different sets of "symbols" (e.g., a geography knowledge base involves entities and relations that may be totally disjoint from those in a biology knowledge base). To bridge this gap, we ground the domain-specific symbols into natural language, and transfer knowledge across domains in the shared realm of natural language.

1.2 Natural Language Interface to Web APIs

Virtual assistants like Apple Siri, Microsoft Cortana, and Amazon Alexa bear the promise of becoming the central natural language interface for personal affairs or smart homes. Towards this goal, a core challenge is *scalability*, i.e., how to extend a virtual assistant to support new services or devices. The current *centralized development* relies on virtual assistant developers themselves and can hardly scale. *Distributed development* is widely believed to be the key to scalability, where third-party service providers or Internet-of-Things (IoT) device manufacturers build

NLIs to their respective services or devices, which can then be easily integrated into virtual assistants. The main technical challenge, however, is how to provide third parties with the capability to build NLIs on their own.

Web API is the standard mean of access to web services in the cloud and IoT devices. For example, through a suite of web APIs called Microsoft Graph, Microsoft provides authenticated access to user data from a wide range of cloud services such as Outlook and Office. With NLI to these APIs, users can enjoy advanced capabilities such as email search (e.g., "*find unread emails from my manager about the Cortana project*") and calendar management (e.g., "*set up a one-hour Skype meeting with John at 10 am tomorrow*"). In a series of collaborations with Microsoft Research and the Cortana team, we developed the *first end-to-end framework for building an NLI to any given web API* [4,11]. The framework comprises a number of innovations, ranging from crowdsourcing pipeline for collecting high-quality, low-cost training data, to interpretable and interactive NLI models.

I also investigated several theoretical aspects of natural language interface. Firstly, based on the compositionality of formal languages, we revealed an *inherent structure* of the NLI problem space, and provided a hierarchical probabilistic model to capture this structure [4]. This structure can be helpful for many tasks. As an example, we have leveraged it to develop an active learning algorithm for crowdsourcing optimization, which lowers the cost for training data collection. Secondly, *interpretability and interactivity* are crucial for a practical NLI, especially in sensitive domains like healthcare and finance. We developed the first interpretable and interactive neural NLI model [11]. The model opens up the black box of neural network by decomposing the prediction into small, interpretable units, each of which is itself a neural network but with pre-defined semantics. We can then explain the model prediction to users in an intuitive way, and solicit fine-grained feedback to correct possible errors. Human subject experiments showed overwhelmingly positive results: Using the interactive NLI, users can successfully complete more tasks while spending less time, leading to a significantly higher level of user satisfaction.

2 Knowledge Harvesting from Massive Texts

Text data is ubiquitous in the digitalized society. For example, up to 80 percent of the information in electronic medical records is in unstructured textual form and therefore largely inaccessible to doctors and biomedical researchers. It is a grand challenge of data science to unlock the rich knowledge buried in massive texts, and provide better management, search, and analytics capabilities for large text corpora. With a focus on text mining and knowledge base construction, my research aims to mine structured, actionable knowledge from massive texts. In particular, I dedicate to developing unsupervised or weakly supervised methods that require minimal human labeling efforts, so that they can be flexibly applied to different application domains. Combining knowledge base construction with natural language interface (§1) enables analysis over raw text data via natural language.

2.1 Knowledge Base Construction with Distant Supervision

Knowledge base construction aims to extract entities and their relationship from texts, e.g., the spouse relationship between *Barack Obama* and *Michelle Obama* from sentence "*Obama and his wife Michelle attended the commencement ceremony at Harvard*." Distant supervision is a widely used principle. By automatically soliciting supervision from external knowledge bases without manual labeling, it bears the promise of providing large amounts of training data for knowledge base construction at scale. However, distant supervision has long suffered from the *wrong labeling problem*: The training labels from distant supervision could be, more often than not, non-representative or even incorrect. Such weak and noisy training data can greatly hurt model performance. We found that global statistics from the entire text corpus and external knowledge base provide a natural way to boost and denoise the training signals from distant supervision [12]. Experiments showed that global statistics can significantly improve state-of-the-art systems for relation extraction, leading to up to 34% reduction in error rate.

2.2 Text Mining in Scientific Literature

The volume of the scientific literature has become prohibitive for researchers and practitioners. For example, over one million new biomedical papers are added into PubMed every year. I have been developing text mining techniques for better management, search, and knowledge discovery in scientific literature. For example, together with two junior PhD students I mentored, we studied how to *categorize scientific publications* [10]. Because of the high dynamics of scientific study, unsupervised methods are highly desired so that the application is not bound by manual labeling costs. We developed the first unsupervised categorization algorithm for scientific publications. This algorithm is a harmonious combination of data mining and deep learning. The key observation is that concepts are the main bearer of category information in documents. For example, a paper likely belongs to the "*deep learning*" category if it contains many "*deep learning*" concepts like "*neural network*," "*max pooling*," and "*activation function*." Therefore, we mine key concepts (significant phrases) in each document with state-of-the-art phrase mining techniques, and then develop a novel cascade embedding approach to categorize the mined concepts. The category of a document can then be robustly determined by jointly considering all of the concepts it contains.

3 Future Research Agenda

My long-term goal is to create *virtual assistants for data analytics* that serve as a unified natural language interface to query, manipulate, and analyze massive and heterogeneous data. With such virtual assistants, users can stay focused on high-level thinking and decision making, instead of overwhelmed by low-level programming and software-specific implementation details. This research direction is still in its infancy, and I am thrilled to continue my research to carry it forward. I am excited to explore the following research opportunities:

1. **Natural Language Interface to Analytics Functions**. Many successful data analytics software and platforms have been developed to meet the pervasive need of data science, which provide analytics functions to view, edit, visualize, and even create advanced predictions on data. However, understanding the large amount of analytics functions and choosing proper ones for the task at hand is still a challenge for non-expert users, especially for complex tasks that require assembling a program of multiple functions. A typical example is as follows. In order to use nocturnal luminosity observed by satellites to study development in China, MIT economist Matt Lowe attempted to use a mainstream geographic information system to analyze geospatial data. The specific analysis he needed is to "*calculate the average luminosity along roads in 1994*." However, he found himself repeatedly puzzling over understanding software-specific geospatial data formats, choosing proper functions among hundreds, and chaining their inputs and outputs. Eventually a complex program with 9 steps was assembled for this seemingly simple and one-off analysis. I envision that an NLI that can automatically transduce user needs into proper analytics function calls will largely bridge this gap. I plan to investigate this new problem in two steps: (1) *Transducing simple user needs into individual function calls*. Analytics functions are similar to APIs, and my experience in NLI to APIs makes me well-prepared for this problem. (2) *Transducing complex user needs into a program of function calls*. With the popularity of machine learning toolkits like Tensorflow and scikit-learn, such NLI may even enable on-demand creation of machine learning models to support user needs like "*tell me the sentiment of these tweets*." I am interested in collaborating with software engineering, program synthesis in particular, researchers to purse this problem.

2. **Theoretical Foundations of Natural Language Interface**. I am particularly interested in answering the following questions: (1) *What is the inherent structure of the NLI problem space?* The inherent structure in a problem can provide strong priors for learning. For example, the manifold assumption, that probability mass mostly lies in a low-dimensional subspace of the original high-dimensional feature space, has been widely used in many computer vision and speech recognition problems to make the learning more efficient. Both natural and formal languages are characterized by *compositionality*, the algebraic capacity to understand and produce a potentially infinite number of novel combinations from known components. If a person understands the meaning of integer "three" and "four" and operator "plus", she can immediately understand and compose

phrases like "three plus four plus four plus three". The lack of compositionality is one main reason for the data inefficiency of neural networks. I am interested in exploring the inherent structure of the NLI problem space based on compositionality. We achieved some preliminary success in [4], but it left more questions than it answered: *For the two kinds of composition respectively from formal and natural languages, how do they interplay? What is the best computational model to capture this structure? Can we develop NLI models with built-in awareness of compositionality?* (2) *How to integrate symbolic and neural computation*. While early NLI models are purely symbolic, now we are shifting towards the other extreme of the spectrum, and study the NLI problem almost exclusively in the realm of neural computing, largely ignoring the symbolic nature of languages. I am interested in the integration of symbolic and neural computation in NLI. One possible direction is to borrow ideas from cognitive science, e.g., Tensor Product Representation from Paul Smolensky that can encode symbolic structures losslessly with numerical tensors.

3. **Mining Structured Knowledge from Massive Texts**. I have particular interests in the following problems: (1) *Combining knowledge base construction and natural language interface*. Organic combination of the two techniques can enable direct natural language analysis over raw text data, which can be applied to many fields including healthcare and social science. For example, we can first construct a knowledge base from electronic medical records, and build an NLI on top of it. Doctors can then directly ask questions to search patient information and make well-informed clinical decisions. (2) *Developing the next-generation search, recommendation, and knowledge discovery system for scientific literature*. Deep text analysis on individual scientific publications and on the entire literature can improve the efficiency of many steps in the scientific workflow. For example, knowledge base construction from scientific literature can bring new opportunities in knowledge discovery and scientific hypothesis generation. For precision medicine, the constructed knowledge base can help answer questions like "*can drug A be repurposed to treat disease B*."

4. **Human-Computer Conversational Interaction**. Decision making is a dynamic process involving multi-round interactions with data for hypothesis generation and verification. Such a process can be naturally accommodated in a conversation between human and computer. Traditional research on dialog and conversational interface, e.g., on tasks like restaurant or flight booking, often assumes fixed "slots" for user needs. For example, the task of restaurant booking has slots like time, party size, etc. The main goal of dialog management is then to interact with the user and fill these slots. However, for the virtual assistant for data analytics, user needs come in unrestricted forms and have no fixed slots. Based on my previous experience in interactive NLI, I plan to study conversational interaction mechanisms in the context of virtual assistant for data analytics. I also look forward to collaborating with researchers in human-computer interaction and cognitive science to study the design of user-friendly interaction mechanisms.

Collaboration. I have taken and will continue interdisciplinary perspectives in my research. I enjoy collaborating with researchers outside computer science to help address their needs of data analytics. For example, I am collaborating with Song Gao from the Department of Geography at University of Wisconsin-Madison (UW-Madison) on natural language interface for geographic databases and information systems. I have built long-term collaboration with researchers from both academic and industrial institutes, including University of Illinois Urbana-Champaign (Jiawei Han, ACM and IEEE fellow), U.S. Army Research Lab (Brian Sadler, IEEE and ARL fellow, senior scientist of U.S. Army), Microsoft Research (Patrick Pantel, Partner Research Manager; Ryen White, Principal Applied Science Manager; Ahmed H. Awadallah, Research Manager), IBM Research (Mudhakar Srivatsa, Principal Research Staff Member), Allen Institute for Artificial Intelligence (Scott Wen-tau Yih, Principal Research Scientist), UW-Madison (Song Gao, Assistant Professor in Department of Geography), among others. In collaboration with Microsoft Research and the Cortana team, our work on NLI to web API has led to a prototype demoed to Bill Gates and a U.S. patent. In collaboration with UIUC, our work on text mining in scientific literature is in the process of transferring to ARL and Department of Defense for technology horizon watch. Together with Xiang Ren from University of Southern California (USC), we presented a tutorial entitled "Construction and Querying of Large-scale Knowledge Bases" in ACM 2017 International Conference on Information and Knowledge Management [9]. With the joining of Pedro Szekely from USC, we will give an updated full-day tutorial in the 27th International World

Wide Web Conference. I am co-organizing the first workshop on "Knowledge Base Construction, Reasoning and Mining," co-located with the 11th ACM International Conference on Web Search and Data Mining in Los Angeles. It is a new effort to bring researchers together to discuss the frontier of knowledge base, featured by a group of world-renowned keynote speakers from diverse fields, including Oren Etzioni (natural language processing), Alon Halevy (database), Monica Lam (system), Christopher Ré (database and machine learning), etc.

Funding. I have extensive experience in funding proposal writing. I wrote the white paper on "Deep Learning for Scientific Taxonomy Construction and Evolution Discovery," which was funded by ARL with $1.5 million in 3 years. During my four years as a lead student researcher in the ARL Network Science Collaborative Technology Alliance, I extensively participated in writing white papers and annual reports, participating in annual evaluations, and visiting allied institutes as representative of UCSB. I also wrote the benchmark and user feedback sections of the NSF funded proposal "Knowledge Graph Query Processing and Benchmarking." I will keep seeking funding opportunities in the future from multiple funding agencies (e.g., NSF, ARL, NIH, and DARPA) and industries.

References

[1] **Yu Su**, Shengqi Yang, Huan Sun, Mudhakar Srivatsa, Sue Kase, Michelle Vanni, Xifeng Yan. Exploiting Relevance Feedback in Knowledge Graph Search. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2015.

[2] **Yu Su**, Huan Sun, Brian Sadler, Mudhakar Srivatsa, Izzeddin Gur, Zenghui Yan, Xifeng Yan. On Generating Characteristic-rich Question Sets for QA Evaluation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2016.

[3] **Yu Su**, Xifeng Yan. Cross-domain Semantic Parsing via Paraphrasing. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2017.

[4] **Yu Su**, Ahmed H. Awadallah, Madian Khabsa, Patrick Pantel, Michael Gamon, Mark Encarnacion. Building Natural Language Interfaces to Web APIs. In *Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM)*, 2017.

[5] Huan Sun, Hao Ma, Xiaodong He, Wen-Tau Yih, **Yu Su**, Xifeng Yan. Table Cell Search for Question Answering. In *Proceedings of the International World Wide Web Conference (WWW)*, 2016.

[6] Semih Yavuz, Izzeddin Gur, **Yu Su**, Xifeng Yan. Improving Semantic Parsing via Answer Type Inference. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2016.

[7] Semih Yavuz, Izzeddin Gur, **Yu Su**, Xifeng Yan. Recovering Question Answering Errors via Query Revision. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2017.

[8] Jie Zhao, **Yu Su**, Ziyu Guan, Huan Sun. An End-to-End Deep Framework for Answer Triggering with a Novel Group-Level Objective. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2017.

[9] Xiang Ren, **Yu Su**, Xifeng Yan. Construction and Querying of Large-scale Knowledge Bases. In *Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM)*, 2017 (tutorial).

[10] Keqian Li, Hanwen Zha, **Yu Su**, Xifeng Yan. Unsupervised Categorization of Scientific Publications. In *Proceedings of the SIAM International Conference on Data Mining (SDM)*, 2018.

[11] **Yu Su**, Ahmed H. Awadallah, Miaosen Wang, Ryen White. Interpretable Natural Language Interface with Modular Sequence-to-Sequence Model. **Submitted** to the *International World Wide Web Conference (WWW)*, 2018.

[12] **Yu Su**, Honglei Liu, Semih Yavuz, Izzeddin Gur, Huan Sun, Xifeng Yan. Global Relation Embedding for Relation Extraction. **Submitted** to the *Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2018 (prepreint arXiv:1704.05958 [cs.CL]).

[13] Izzeddin Gur, Semih Yavuz, **Yu Su**, Xifeng Yan. Harnessing Predictive Uncertainty for Understanding and Regularizing Neural Predictions. **Submitted** to the *Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2018.

[14] Li Dong, Jonathan Mallinson, Siva Reddy, and Mirella Lapata. Learning to paraphrase for question answering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2017.