Research Statement
Steven Bethard

Natural language processing (NLP) is the science of teaching computers to understand human language. It is a critical area of research for supporting a wide range of applications, from patient-centered research to intelligence analysis to measuring student reading comprehension. I contribute to this field through research in the area of *information extraction*, a subfield of NLP that focuses on how to translate text into a structured form that computers can more easily search and analyze. My vision of information extraction research has to date brought in more than $10,000,000 of funding from the National Institutes of Health, Department of Defense, and National Science Foundation.

I am best known for my research on *timeline extraction*: how the events and times of a narrative can be extracted and integrated into a single coherent representation of the temporal backbone of the narrative. My articles in this field have more than 1000 citations, and my research systems are top-ranked in worldwide competitions of timeline extraction algorithms (Bethard and Martin 2007; Bethard 2013b). I am also deeply involved in research on NLP algorithms for *medical applications*, where in addition to my own research, I have regularly engaged the research community both by organizing the annual Clinical NLP workshop and by hosting international shared tasks where research groups around the world compete to solve clinical information extraction problems (Clinical TempEval in 2015-2017, Parsing Time Normalizations in 2018, and Source Free Domain Adaptation in 2021).

The following sections detail my main research directions.

Extracting timelines from text

I have an extensive record of research on machine-learning approaches to the extraction of timelines from unstructured data. The goal of such research is to teach computers to understand the language of time, so that they can, for example, convert expressions like *the day before yesterday* or *three weeks postoperative* to machine-readable forms like 2019-05-15. This is a challenging task due to the often vague or ambiguous nature of how we speak and write about time. Nonetheless, I have shown that using linguistic knowledge to constrain timeline extraction models yields more precise predictions (Bethard, Martin, and Klingenstein 2007; Bethard 2013b); that modeling timeline extraction as translation from text to semantic representations results in more accurate and expressive models (Bethard 2013a; Laparra, Xu, and Bethard 2018); and that the flexibility of neural network models enables a variety of new and powerful algorithms for timeline extraction (Xu, Laparra, et al. 2019; Miller, Bethard, et al. 2023).

Linking text to medical and geospatial ontologies

My research has also investigated algorithms for linking phrases in text to their standard forms in medical and geospatial ontologies. Such algorithms must solve both the problem that many phrases may refer to the same concept (e.g., "myocardial infarction" and "heart attack") and the problem that a single phrase may refer to different concepts depending on the context (e.g., "Paris" could mean the one in France, the one in Texas, etc.). I have demonstrated that deep neural networks can learn to solve both of these problems more successfully than prior approaches (Xu, Zhang, et al. 2020), that they can be pre-trained directly on ontologies to more deeply integrate knowledge into

the models (Xu and Bethard 2021), and that they are adept at incorporating the context necessary to disambiguate terms (Zhang and Bethard 2023).

Adapting machine learning models to new domains

My research on domain adaptation has demonstrated that big unlabeled data can be combined with small labeled data to make machine-learning models more robust to changes in domains. I have shown that unsupervised methods can learn correspondences between domains (Sapkota, Solorio, et al. 2016), and that pre-training neural network models improves the linguistic capacity of such models (Lin, Bethard, et al. 2020). I am also a pioneer of the new field of source-free domain adaptation, where trained models must be adapted without access to the original training data, a common scenario with medical institutions (Laparra, Bethard, et al. 2020).

Encouraging replicable research

Throughout my career, I have recognized the need for replicable solutions to computational problems. I have helped develop programming frameworks for natural language processing including ClearTK, UIMA, cTAKES, Stanford CoreNLP, and NLTK (Manning, Surdeanu, et al. 2014; Bethard, Ogren, et al. 2014). These frameworks have tens of thousands of users ranging from academia to industry. I have also worked to encourage the direct comparison of different research approaches through the organization of international shared tasks for natural language processing problems (Bethard, Derczynski, et al. 2015; Bethard, Savova, Chen, et al. 2016; Bethard, Savova, Palmer, et al. 2017; Laparra, Xu, Elsayed, et al. 2018; Laparra, Su, et al. 2021). These shared tasks have helped establish the state of the art in areas such as timeline extraction and domain adaptation.

Collaborating across disciplines

I have secured more than $10,000,000 in grants from the National Institutes of Health, Department of Defense, and National Science Foundation. All of these grants involve interdisciplinary teams, often spread across multiple institutions, where I serve in the role of the expert on how to extract timelines, locations, medical concepts, etc. from text. In my medical NLP research, I have engaged with doctors and medical researchers from institutions such as Boston Children's Hospital, Columbia University, University of Miami, and the University of Colorado to obtain funding for our research from the National Institutes of Health. My temporal and geospatial research has been in collaboration with experts in causal modeling and environmental policy from institutions such as Harvard University and the University of Arizona, where we have been successful in funding our research through the Department of Defense and the National Science Foundation. All these funded collaborations have enabled me to maintain a vibrant lab of students and postdocs who actively support my research agenda.