THEORYON: A DESIGN FRAMEWORK AND SYSTEM FOR UNLOCKING BEHAVIORAL KNOWLEDGE THROUGH ONTOLOGY LEARNING

Jingjing Li
Assistant Professor of Information Technology
McIntire School of Commerce, University of Virginia,
Charlottesville, VA 22903 U.S.A. {jl9rf@comm.virginia.edu}

Kai Larsen
Associate Professor of Information Management
Leeds School of Business, University of Colorado,
Boulder, CO 80309 U.S.A. {Kai.Larsen@colorado.edu}

Ahmed Abbasi
Joe and Jane Giovanini Professor of IT, Analytics, and Operations
Mendoza College of Business, University of Notre Dame
Notre Dame, IN 46556 U.S.A {aabbasi@nd.edu}

ABSTRACT

The scholarly information-seeking process for behavioral research consists of three phases: search, access, and processing of past research. Existing IT artifacts, such as Google Scholar, have in part addressed the search and access phases, but fall short of facilitating the processing phase, creating a knowledge inaccessibility problem. We propose a behavioral ontology learning from text (BOLT) design framework that presents concrete prescriptions for developing systems capable of supporting researchers during their processing of behavioral knowledge. Based upon BOLT, we developed a search engine—TheoryOn—to allow researchers to directly search for *constructs, construct relationships*, *antecedents,* and *consequents,* and to easily integrate related *theories*. Our design framework and search engine were rigorously evaluated through a series of data mining experiments, a randomized user experiment, and an applicability check. The data mining experiment results lent credence to the design principles prescribed by BOLT. The randomized experiment compared TheoryOn with EBSCOhost and Google Scholar across four information retrieval tasks, illustrating TheoryOn's ability to reduce false positives and false negatives during the information-seeking process. Furthermore, an in-depth applicability check with IS scholars offered qualitative support for the efficacy of an ontology-based search and the usefulness of TheoryOn during the processing phase of existing research. The evaluation results collectively underscore the significance of our proposed design artifacts for addressing the knowledge inaccessibility problem for behavioral research literature.

Keywords: Behavioral ontology learning design framework, design science research, text analytics, machine learning, randomized experiment, applicability check

INTRODUCTION

Behavioral researchers continually search for and develop theories to improve disciplinary understanding of key phenomena. For example, the theory of planned behaviors that explains an individual's intention to engage in a certain behavior has received more than 70,000 citations (Ajzen 1991). Hundreds of theories have been developed or extended (Soper and Turel 2015) to facilitate the understanding of real-world information systems phenomena, some receiving tens of thousands of citations (e.g., Davis 1989; Venkatesh et al. 2003). Paradoxically, though, the rich academic literature on human behavior has become expansive to the point of incognizance over the past few decades (Weber 2012). Since behavioral research takes a concept-centric perspective, the completeness of any literature search is often defined as the proportion of relevant constructs retrieved (Webster and Watson 2002). In this regard, studies have shown that researchers remain largely unaware of the majority of research, especially outside of their own disciplines, but also within narrow research areas (Colquitt and Zapata-Phelan 2007). Larsen and Bong (2016) have shown that even for a small set of full-text articles, experts could retrieve, on average, fewer than 10% of the articles that would be valuable for a literature review and knowledge acquisition.

The result is knowledge inaccessibility in behavioral research, here defined as the situation that behavioral knowledge embedded in the extant large-scale literature may not be accessed by researchers in a comprehensive and accurate manner. Knowledge inaccessibility could have a considerable negative impact on behavioral research in at least four ways. First, with knowledge inaccessibility, researchers are prone to literature fragmentation and end up reinventing constructs, relationships, or hypotheses already introduced by others. This can result in wasted and redundant research efforts (Spell 2001), possible errors such as spurious gapspotting and gap-patching, and the generation of marginal research (Rai 2017). Second, it prevents the building of cumulative traditions in which researchers build on each other's previous work and that "definitions, topics, and concepts are shared" (Keen 1980), thus threatening the development and progression of a research field (Im and Straub 2012). Third, it introduces inefficiencies in research processes and knowledge acquisition and construction, leaving the research community to be slow in accommodating emerging contexts (Quirchmayer et al. 2012), low in research topic agility (Peffers 2002), and vulnerable to rapid environmental change (Trinh-Phuong et al. 2012). Finally, as behavioral research spans multiple disciplines, including medicine, psychology, sociology, education, and economics, impediments to the knowledge creation process and spurious research findings resulting from knowledge inaccessibility may exact tremendous monetary and social costs (Weber 2012).

The research on information-seeking behaviors in behavioral research could shed light on how knowledge inaccessibility arises. Unlike natural sciences whose theories are consist of strictly universal statements and languages (Popper 1980), behavioral theories often measure beliefs, expectations, attitudes, and emotions through constructs and relationships defined by malleable and ever-changing language systems (Arnulf et al. 2014; 2018; Larsen et al. 2013). Hence, it is important to adopt a construct-centric view (Webster and Watson 2002) and clarify and synthesize construct relationships during the scholarly information search process.

Specifically, behavioral researchers' information seeking can be categorized into phases, including *searching*, *accessing*, and *processing* (Meho and Tibbo 2003), of which the *accessing* phase serves as a conduit between the critical *searching* and *processing* phases. Existing IT artifacts, such as full-text search engines, are well suited for the *searching* phase, in which the process of identifying relevant and potentially relevant materials is initiated. For instance, Google Scholar and EBSCOhost provide keyword searches of the free text in abstracts or full texts and incorporate article-level citation analysis and usage statistics for results ranking (Beel and Gipp 2010). However, the majority of knowledge inaccessibility issues manifest in the *processing* phase, which entails extraction, synthesis, and analysis of concepts across articles. High false-positive rates in full-text search engines due to lack of behavioral knowledge extraction can mislead researchers into prematurely ending the information-seeking process (Boeker et al. 2013). False negatives, demonstrated as confirmation biases (White 2013), could also hinder the completeness of the *processing* outcomes. Specifically, confirmation biases occur as a result of individual researchers' and research fields' proclivity towards "unwitting selectivity in the acquisition and use of evidence" (Nickerson 1998, p. 175) and are amplified by full-text search engines' keyword matching and data-dependent ranking algorithms. In this sense, complementing full-text search engines with new search artifacts capable of disembedding behavioral knowledge to better support the *processing* phase may help enhance information-seeking abilities.

To alleviate the knowledge inaccessibility problem, this paper proposes two design artifacts: a behavioral ontology learning from text (*BOLT*) design framework and an ontology-based search engine *TheoryOn*, to disembed behavioral knowledge from existing, large-scale publications. Using relevant behavioral research (Baron and Kenny 1986; Larsen and Bong 2016; Larsen et al. 2019; Weber 2012), our BOLT design framework views behavioral knowledge as a specialized type of ontology whose core parts include *hypotheses*, *constructs*, and *construct relationships*. Hence, effective behavioral ontology learning entails appropriate extraction of these parts. Referring to ontology learning from text literature (e.g., Wong et al. 2012) and the pertinent natural language processing (NLP) research, BOLT identifies the underlying tasks and prescribes the best potential techniques. We further used the proposed design framework to develop the ontology-based search engine TheoryOn, which allows researchers to directly search for *constructs*, *construct relationships*, and *theoretically related constructs*, as well as to easily integrate *related theories*. We also conducted a multifaceted evaluation of TheoryOn (Gill and Hevner 2013; Hevner et al. 2004) which included ontology learning method and system comparison experiments, a randomized user experiment comparing it with the EBSCOhost and Google Scholar search engines, and an applicability check. Overall, the contribution of our work represents an instance of *exaptation* in which we adapted solutions from the ontology learning field to a new problem, that of disembedding behavioral knowledge from large-scale behavioral publications (Gregor and Hevner 2013).

BACKGROUND: LIMITATIONS OF EXISTING SEARCH ENGINES TO SUPPORT SCHOLARLY LITERATURE REVIEW

At a high level, behavioral researchers' information seeking can be categorized into three closely inter-related phases: *searching*, *accessing*, and *processing* (Ellis 1989; Meho and Tibbo 2003). *Searching* encompasses "the period where identifying relevant and potentially relevant materials is initiated" (Meho and Tibbo 2003, p. 584). This phase includes steps such as initial search, following chains of citations, and casually browsing selected articles (Ellis 1989). *Processing* is where synthesizing and analyzing across articles and concepts takes place (Meho and Tibbo 2003). This phase is especially important for behavioral research because, unlike natural sciences, behavioral research measures beliefs, perceptions, and emotions that are less amenable to describe in universal languages. Hence, there is a greater need to scrutinize, differentiate, filter, organize, and amalgamate information across articles. Since information seeking in behavioral research is a non-linear process, the *accessing* phase simply serves as a conduit between the critical *searching* and *processing* phases. These phases are consistent with information-seeking stages identified through our survey of IS scholars (see Appendix D for details).

Full-text academic search engines such as Google Scholar and EBSCOhost are especially well suited to supporting the *searching* phase. They allow individual users to specify search queries that represent their search needs and return as search results as a subset of articles that contain all/some keywords from the search query (Beel and Gipp 2010). They also allow researchers to conduct keyword searches within articles that cited a relevant paper. The efficiency and ease of use of full-text search engines are ideal for the initial search phase, high-level browsing of potentially relevant articles, and quickly making sense of an area of research through query keyword expansion and citation network traversal. Conversely, full-text engines are not as well suited to supporting the *processing* phase of information-seeking behavior, where false positive and false negative errors can adversely affect synthesizing and analyzing activities.

First, full-text search engines do not extract behavioral knowledge-relevant metadata embedded in articles (e.g., constructs and construct relationships), which can lead to a large number of false positives in behavioral knowledge searches. For instance, a search for the construct *perceived usefulness*, intended to represent the perceived belief that a system can enhance job performance (Davis 1989), returned 90,200 results in Google Scholar (retrieved on 1/31/2019). Rather than the actual construct, *perceived usefulness*, most of the returned articles contained the loosely used phrase, *perceived usefulness*, or constructs carrying the same name but representing different latent concepts, such as Nelson's (1991) *perceived usefulness* scale, which measures the perceived importance of skill proficiency on job performance. Boeker et al. (2013) found that across 14 existing systematic studies, the precision of Google Scholar was

0.13%. Similarly, Yousafzai et al. (2007) evaluated 36,463 articles in a Google Scholar search result for the Technology Acceptance Model and found precision to be 0.39%, indicating that finding all relevant articles would require evaluating hundreds of false positives for each truly relevant article found.

Second, keyword matching and citation- and usage-based ranking, although efficient and effective in supporting the *searching* phase, may lead to heavy false negatives during the processing phase (Larsen et al. 2019). The false negatives are largely caused by researcher-level and field-level confirmation biases. Search queries based on keywords confirm researchers' preexisting beliefs about construct names and research topics. Correspondingly, a search, for example, for the *subjective norm* construct may altogether miss articles that employ identical operationalizations, but with different names, such as *social factors* or *image*. This type of confirmation bias, originating from researchers' tendency to confirm existing beliefs while neglecting viable alternatives, is referred to as researcher-level confirmation bias (Bushman and Wells 2001). Similarly, the citation-network- and usage-statistics-based ordering of results (Beel and Gipp 2010), typical of search engines such as Google Scholar, introduces a field-level confirmation bias. Due to researcher-level confirmation bias, strongly confirmative articles receive more clicks, downloads, and citations (White 2013), further promoting these articles to the top of the search results (Beel and Gipp 2010) and resulting in a field-level confirmation bias.

Admittedly, researchers and practitioners from scholarly information retrieval fields have proposed academic support IT artifacts that go beyond keyword-based indexing and allow direct search on metadata from academic articles. For example, Quan et al. (2004) applied a Fuzzy Formal Concept Analysis (FFCA) method to build a scholarly ontology from a citation database, an important step toward building an ontology-based search engine. Semanticscholar.org extracts authors, journals, conferences, figures, references, and topics from academic articles to facilitate a more nuanced search. Similarly, Microsoft Academic Search extracts seven entity types, including authors, affiliations, title, year, journal, conference series, and field of study to help users quickly process knowledge embedded in academic articles. However, these systems focus on metadata such as authors, citations, and journals, and do not incorporate provisions for behavioral knowledge disembedding (e.g., hypotheses, constructs, and relationships), rendering them less effective for facilitating the processing phase of scholarly information-seeking and solving the behavioral knowledge inaccessibility problem.

Another related scholarly information retrieval field is biomedical text mining, which utilizes NLP techniques to extract genes, proteins, drugs, diseases, and their relations from the biomedical literature (e.g., Luo et al. 2016). However, biologists have the advantage of gene nomenclature committees (Eyre et al. 2006) and the good fortune of working with constructs—genes—that have more precise definitions and are more amenable to consistent measurement through commonly accepted instruments. Conversely, behavioral constructs are often defined by malleable and ever-changing language systems. Consequently, the methods suitable for biomedical text mining may be less so in behavioral knowledge disembedding.

DESIGN FRAMEWORK FOR DISEMBEDDING BEHAVIORAL KNOWLEDGE

Given the need for disembedding behavioral knowledge to improve the scholarly information-seeking process and the lack of dedicated IT artifacts that address this need, an important set of questions arises. How do we define behavioral knowledge? What are the key features and capabilities that behavioral knowledge disembedding systems should support? What are the necessary tasks to accomplish them? Relative to keyword-based search engines, metadata systems, and biomedical text analytics tools, systems geared towards extracting behavioral knowledge face greater ambiguity. Behavioral research involves various philosophical paradigms, spans numerous disciplines, and employs a number of research methods, which lead to a considerable level of disagreement about what constitutes behavioral knowledge (e.g., Corley and Gioia 2011; Gregor 2006; Weber 2012). Consequently, there is lack of clarity on the key features and capabilities that should be supported by behavioral knowledge disembedding systems, not to mention the necessary tasks and techniques to accomplish them.

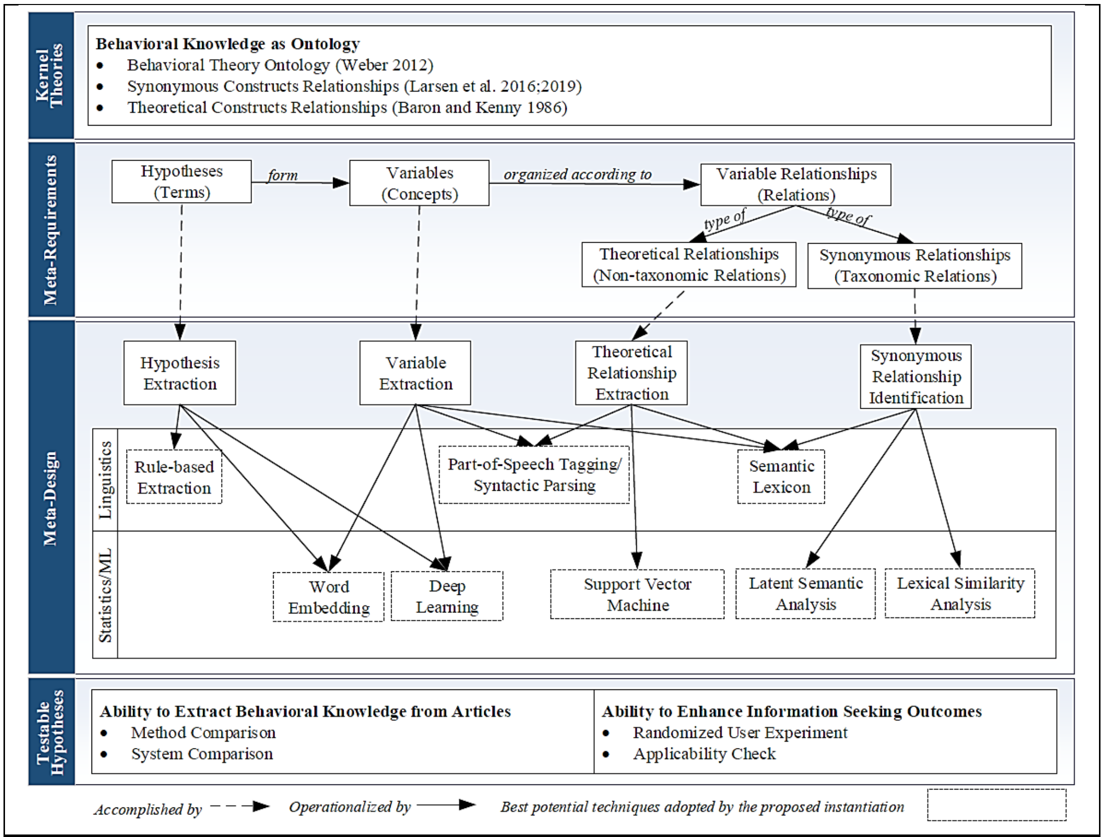

Kernel Theories
Behavioral Knowledge as Ontology
Behavioral Theory Ontology (Weber 2012)
Synonymous Constructs Relationships (Larsen et al. 2016;2019)
Theoretical Constructs Relationships (Baron and Kenny 1986)
Meta-Requirements
Hypotheses (Terms)
form
Variables (Concepts)
organized according to
Variable Relationships (Relations)
type of
type of
Theoretical Relationships (Non-taxonomic Relations)
Synonymous Relationships (Taxonomic Relations)
Meta-Design
Hypothesis Extraction
Variable Extraction
Theoretical Relationship Extraction
Synonymous Relationship Identification
Linguistics
Rule-based Extraction
Part-of-Speech Tagging/ Syntactic Parsing
Semantic Lexicon
Statistics/ML
Word Embedding
Deep Learning
Support Vector Machine
Latent Semantic Analysis
Lexical Similarity Analysis
Testable Hypotheses
Ability to Extract Behavioral Knowledge from Articles
Method Comparison
System Comparison
Ability to Enhance Information Seeking Outcomes
Randomized User Experiment
Applicability Check
Accomplished by
Operationalized by
Best potential techniques adopted by the proposed instantiation

Figure 1. Behavior Ontology Learning from Text (BOLT) Framework

Design guidelines are needed due to this complexity of properly representing behavioral knowledge and the resulting ambiguity regarding the key features, capabilities, and tasks a behavioral knowledge disembedding system should support. We therefore propose a design framework named behavioral ontology learning from text (BOLT), to guide the development of systems for *extracting behavioral knowledge encompassed in large-scale, multidisciplinary publication databases*. According to the design science paradigm (Hevner et al. 2004; Walls et al. 1992), design is both a product and a process. The design product concerns a set of requirements and design characteristics to guide IT artifact construction. Meanwhile, the design process involves steps and procedures to construct the IT artifact, and typically follows a highly iterative process consisting of building and evaluating (March and Simon 1995). Our design framework focuses on the design product, which is composed of *kernel theories*, *meta-requirements*, *meta-design*, and *testable hypotheses* (Walls et al. 1992; Abbasi and Chen 2008). Figure 1 depicts our BOLT design framework.

According to Walls et al. (1992), kernel theories are derived from the natural and social sciences and are used to govern meta-requirements. However, as noted by Arazy et al. (2010), theories from those domains are rarely used as-is because their scope and granularity are often inadequate for a specific design problem. Hence, we draw on multiple behavioral studies (Baron and Kenny 1986; Larsen and Bong 2016; Larsen et al. 2019; Weber 2012) as *kernel theories* to identify what constitutes behavioral knowledge. Specifically, behavioral knowledge can be considered as theories encompassing originating, extending, and subscribing behavioral articles (theory instances), and each of these articles is a specialized ontology whose core parts include *constructs* and their *relationships*. Accordingly, the *meta-requirements* frame behavioral knowledge disembedding as a specific ontology learning process (Buitelaar et al. 2005) which needs to support the extraction of *hypotheses (terms)*, *variables (concepts)*, *theoretical relationships (non-taxonomic relations)*, and *synonymous relationships (taxonomic relations)*[1]

[1] Synonymous relationships depict whether two or more variables are referring to the same underlying meaning, which could be used to build taxonomic relationships. For example, *performance expectancy* in Venkatesh et al. (2003) and *perceived usefulness* in Venkatesh and Morris (2000) are both measuring an individual's perception of a

from behavioral articles. The *meta-design* identifies the underlying tasks and provides viable techniques. Four BOLT tasks are identified as a result: *hypothesis extraction, variable extraction, theoretical relationship extraction,* and *synonymous relationship identification*. Based on the ontology learning and pertinent NLP literature, we organize viable techniques into two categories—*linguistics* and *statistics/machine learning (ML)*. For each BOLT task, a thoughtful selection and coordination of techniques across these two categories are needed. In this study, we present the best potential techniques (as of the time of publication) as an implementation example. Interested researchers can replace them with better techniques when linguistics and statistics/ML methods advance in the future. Finally, we propose *testable hypotheses* to empirically evaluate how well the proposed meta-design meets the meta-requirements. These hypotheses involve both the ability to extract behavioral knowledge from behavioral articles and the ability to enhance behavioral researchers' information-seeking outcomes. To test them, a multifaceted evaluation solution is needed which includes method and system comparisons, randomized user experiments, and applicability checks. In the following section, we discuss the four components of the BOLT framework in detail.

Kernel Theories

As noted, there is lack of consensus on how to best define behavioral knowledge and its key components. Weber (2012) argues that theoretical development is a central behavioral research endeavor. Hence, theories could represent the most important type of behavioral knowledge. Larsen et al. (2019) suggest that a theory consists of a set of publications, including the originating publication, the most influential extensions of the original article, and all theory-

system's usefulness, despite of different words. Hence, they can be considered as hyponyms of the more general construct perceived usefulness.

subscribing articles. Collectively, these articles are referred to as *theory instances*. According to Weber (2012, p.3), a theory instance is "a particular kind of model that is intended to account for some subset of phenomena in the real world." [2] Specifically, the subset of phenomena usually pertains to classes of things in a domain, and the model is an abstracted, simplified, concise representation that explains and predicts a phenomenon (Parsons and Wand 2013). In this light, theory instances "can be conceived as *specialized* ontologies—instances of [Bunge's (1977; 1979)[3]] *general* ontology (a theory about the nature of and makeup of the real world, in general)" (Weber 2012, p. 3).

The core parts of a theory instance include *constructs*, their *relationships*, and the *state* they cover (Weber 2012). In behavioral research, a construct represents "an *attribute in general* of some *class of things* in its domain" (Weber 2012, p. 7). Constructs serve a central role in a theory instance because their definition directly governs the meaning of construct relationships and the state space of a theory. Construct relationships can be in the form of correlation, causation, or synonymous relationships. Correlation or causation, referred to as theoretical relationships hereafter, can be categorized as *main effect*, *moderation*, and *mediation* (Baron and Kenny 1986). Main effect pertains to a direct theoretical relationship between two constructs, moderation involves a third construct affecting the strength or direction of a theoretical relationship, and mediation entails an intermediate construct between two theoretically related constructs. In contrast, a synonymous relationship represents an "is-a" association between

[2] Weber's (2012) notion of theory is best aligned with Gregor's (2006) Type IV theory—a theory for explanation and prediction.

[3] There may be alternative notions about the mapping between behavioral theories and ontologies, but such is not the focus of this paper. The adoption of Weber's theory allows many of the ontology learning tasks and techniques to be adapted to guide extraction of behavioral knowledge from large databases of behavioral publications.

different constructs, within or across articles, referring to the same underlying meaning (Larsen and Bong 2016). For example, in Venkatesh and Morris (2000), the two construct mentions of *behavioral intention* in hypotheses H1 and H2a refer to the same meaning; and the construct *performance expectancy* in Venkatesh (2003) is synonymous with that of *perceived usefulness* proposed by Davis (1989). Finally, the state of a theory instance is a conceivable state space that falls within a theory's boundary, which is determined by "the range of values that each construct in the theory might cover" (Weber 2012, p. 11). Taken together, this ontology-centric view of behavioral knowledge afforded by the amalgamation of the aforementioned kernel theories provides an appropriate mechanism for disembedding behavioral knowledge by underscoring the importance of focusing on the core parts embodied in theory instances.

Meta-Requirements

As noted, behavioral knowledge could be considered to be comprised of theories encompassing multiple theory instances, each of which can be conceived as a specialized Bunge's (1977, 1979) ontology. The core parts of each theory instance include constructs, their relationships, and the state they cover. "The parts of a theory need to be described precisely because they circumscribe the boundary or domain of the theory – that is, the phenomena it is intended to cover" (Weber 2012, p. 6). Therefore, the effective disembedding of behavioral knowledge calls for behavioral ontology learning capable of extracting those parts.

Ontology learning pertains to the development and use of various automated techniques to extract the key components of ontology from large-scale textual data (Buitelaar et al. 2005). The goal for ontology learning from text is to bridge the gap in a data context that "scores highest on availability and lowest on accessibility" (Biemann 2005, p. 79)—an objective that nicely parallels our behavioral knowledge inaccessibility alleviation objective. Buitelaar et al. (2005)

synthesized the ontology learning literature into a core set of five “layer cake” outputs: *terms, concepts*, *taxonomic relations*, *non-taxonomic relations*, and *axioms*. The outputs above are ordered, meaning that each output is a prerequisite for obtaining the next.

- *Terms* are lexical components that contain important pieces for an ontology.
- *Concepts* are formed by leveraging terms to represent objects.
- *Taxonomic and non-taxonomic relations* depict relationships between concepts. *Taxonomic relations* are focused on extracting “is–a” relations (hypernym/hyponym). An example would be “a duck is a type of the concept waterfowl.” *Non-taxonomic relations* are non-hierarchical relations. For instance, “a concept duck is often ‘found near’ ponds.”
- *Axioms* are rules defined over concepts.

In the behavioral ontology learning context, *variables*[4] represent general attributes of some class of things covering a phenomenon and are best aligned with *concepts*. *Synonymous relationships* depict whether two or more variables are referring to the same underlying meaning. These relationships can be used to build a construct hierarchy. Hence, they can be mapped to *taxonomic relations*. *Theoretical relationships* representing correlation and causation are best related to *non-taxonomic relations*. An important starting point in ontology learning, however, is to identify lexical components that encompass variables and relationships. Fortunately, in behavioral research, an article belonging to Weber’s (2012) notion of theory—theory for explanation and prediction—usually presents the behavioral theory through hypotheses, which

[4] We use the term “variable” to encompass constructs as well as non-construct variables, which play a key role in the theory and warrant extraction (e.g., demographics).

include a statement describing the relationships between variables.[5] Hence, we consider *hypotheses* as the *terms* of an ontology learning layer cake.

In summary, behavioral ontology learning entails four ordered layer cake outputs: *hypotheses (terms)*, *constructs (concepts)*, *theoretical relationships (non-taxonomic relations)*, and *synonymous relationships (taxonomic relations)*. We acknowledge that other behavioral ontology parts, such as the state a theory covers, are also important to disembed (Weber et al. 2012; Schryen et al. 2017); however, we leave these for future research.

Meta-Design

Based on the aforementioned meta-requirements, the underlying tasks for disembedding behavioral knowledge are *hypothesis extraction*, *variable extraction*, *theoretical relationships extraction,* and *synonymous relationships identification*. A plethora of ontology learning techniques can be leveraged to support these tasks, two of which are pertinent to our context: linguistics and statistics (Wong et al. 2012). *Linguistics* techniques extract linguistic features such as parts of speech (POS), syntactical structure analysis, and dependency analysis. When texts follow a prescribed linguistic pattern, linguistic rules can be derived to extract relevant lexical components for ontology building. *Statistics* techniques are primarily derived from information retrieval, machine learning (ML), and data mining literature. These types of techniques usually rely on linguistic components as underlying input features. Sample techniques include co-occurrence analysis, latent semantic analysis, clustering, and association rules. However, ML approaches, such as classification, sequence labeling techniques, deep learning,

[5] We randomly sampled 40 articles from *MIS Quarterly* and the *Journal of Applied Psychology* and had domain experts place the contained hypotheses into two classes: "supported" and "unsupported." The results showed that hypotheses were unsupported 23.7% to 31% of the time in the respective publications. Unsupported hypotheses are important drivers of theoretical progress (Popper 1959). Indeed, in meta-analysis, supported and unsupported relationships are equally important.

and support vector machines (SVMs), that are popular in the state-of-the-art text mining literature for concept and concept relation extraction have been underutilized in the ontology learning domain (Asim et al. 2018; Wong et al. 2012). To better emphasize the immense potential of ML methods such as deep learning, we expand the statistics techniques category in our framework to be *statistics/ML* techniques. In the following section, we describe each BOLT task and identify the best potential supporting techniques from the linguistics and statistics/ML categories. The alternative techniques are described in Appendix A.

Hypothesis Extraction

The initial task in BOLT is *hypothesis extraction*. Hypothesis formats for several behavioral disciplines (e.g., IS, management, marketing) are formalized. For example, the following shows a typical hypothesis from a behavioral research article by Venkatesh and Morris (2000), which stands as an independent paragraph:

H1: Perceived usefulness will influence behavioral intention to use a system more strongly for men than it will influence women.

A formatting rule could be derived from the above hypothesis: "a capitalized H + a number [+a letter] + a colon + a capitalized word + a string of words + a period." Similarly, additional rules could be generated by enumerating a sufficient number of articles. Per the ontology learning literature, when the hypothesis text elements have apparent linguistic cues, *rule-based extraction* may be best suited for extracting morphological patterns.

However, hypotheses in many other behavioral research areas may not follow the traditional format identified above. In such situations, *statistics/ML* techniques, such as text classification, can be used to discover appropriate patterns. These methods use training data sets to build text classifiers for automatically predicting class label (i.e., whether a sentence is a hypothesis) based on extensive manual feature engineering. Examples of baseline supervised statistical methods include maximum entropy (Berger et al. 1996) and naïve Bayes (Ng and Jordan 2002). The recent development in ML affords opportunities to further enhance the classification performance (Cho et al. 2014). Examples include *word embedding* (Mikolov et al. 2013) for automated feature engineering and *deep learning* such as convolutional neural networks (CNN) (LeCun et al. 1998) and recurrent neural networks (RNN) (Mikolov et al. 2010) for modeling non-linear, complex patterns. Word embedding encapsulates word-level distributional semantics. Word2Vec (Mikolov et al. 2013) and GloVe (Pennington et al. 2014) are two common methods for deriving word embeddings. The former uses a neural network model to represent a word's distributional semantics by examining its surrounding words, and the latter uses a matrix decomposition method over a large-scale data set to model similarities between words. The CNN method utilizes convolutional filters to learn character-level morphological patterns or local features that are critical to differentiating classes of texts (e.g., Kim 2014). The RNN method models long-distance dependency between linguistic components (Hochreiter and Schmidhuber 1997). Consequently, methods that combine rule-based techniques for capturing common templates with state-of-the-art ML to account for variances in hypothesis articulation and formatting across authors and disciplines in behavioral research may be well suited for hypothesis extraction.

Variable Extraction

The second step in the BOLT framework involves deriving variables from the extracted hypotheses. The following is an example showing variables contained in the first hypothesis of Venkatesh and Morris (2000). Inside, Outside, Beginning (IOB) tagging is commonly used for concept/entity extraction, where "B" marks the beginning word of a variable, "I" labels the words inside a variable, and "O" represents the words outside a variable. In the following hypothesis, perceived usefulness and behavioral intention are constructs, and men and women are the values for the gender variable.

H1/O :/O **Perceived/B usefulness/I** will/O influence/O **behavioral/B intention/I** to/O use/O a/O system/O more/O strongly/O for/O **men/B** than/O it/O will/O for/O **women/B**./O

A survey of the existing techniques addressing concept or entity extraction (similar to variable extraction) reveals that *statistics/ML* methods are highly applicable (Hobbs and Riloff 2010). Specifically, supervised ML methods convert variable extraction into a sequence labeling problem, where each word in a hypothesis has a specific class label (e.g., the IOB tags), and the dependency among the class labels is explicitly considered (Lafferty et al. 2001). Hence, the sequence labeling task is to predict the most probable class label for each word, depending on its linguistic features as well as its relationship to the surrounding words with their linguistic features. The status quo *statistics/ML* methods for sequence labeling problems are feature-based methods, such as the hidden markov model (HMM) (Rabiner 1989) and conditional random fields (CRF) (Lafferty et al. 2001), whose performances are heavily dependent upon labor-intensive feature engineering.

The recent advances in ML present opportunities for enhancing performance by involving a hybridized deep learning method (Yadav and Bethard, 2018). Such methods could make use of character-level CNN to incorporate morphological patterns (e.g., capitalization in words) and bidirectional long short-term memory (Bi-LSTM) (Hochreiter and Schmidhuber 1997), a specific type of RNN, to characterize the dependency among linguistic inputs. In addition, incorporating a CRF layer in a deep learning architecture may further enhance performance by considering the dependency among class labels (e.g., Ma and Hovey 2016).

Additionally, incorporating word embedding that reflects linguistic richness and domain knowledge could also help (Huang et al. 2015). For example, certain types of words and phrases are known to be more likely to appear in variables (e.g., noun phrases starting with the word "perceived"). This information could be leveraged in customized word embedding of a deep learning architecture, even with limited training data (Huang et al. 2015). In summary, methods at the intersection of customized word embedding, domain-adapted features/lexicons, and hybridized deep learning architectures may be well suited for variable extraction.

Theoretical Relationship Extraction

Once variables embedded in hypotheses are extracted, the third task involves identifying relationships among them. The following examples show three hypotheses annotated with extracted variables representing a main, a moderation, and a mediation relationship, from Krosgaard et al. (2002). Specifically, managerial trustworthy behavior is an antecedent (AT), and employees' trust in the manager is a consequent (CT), perceived fairness of human resource policies is a moderator (MOD), and employees' attributions of responsibility for an event are mediators (MED).

Hypothesis 3 *(main)*: [Managerial trustworthy behavior]$_{\text{AT}}$, in the form of communication and concern, is positively related to [employees' trust in the manager]$_{\text{CT}}$.

Hypothesis 4 (*moderation*): The relationship between [managerial trustworthy behavior]$_{\text{AT}}$ and [employees' trust in the manager]$_{\text{CT}}$ is moderated by the [perceived fairness of human resource policies]$_{\text{MOD}}$.

Hypothesis 5 (*mediation*): The relationship between [managerial trustworthy behavior]$_{\text{AT}}$ and [employees' trust in the manager]$_{\text{CT}}$ is mediated by [employees' attributions of responsibility for the event]$_{\text{MED}}$.

To identify such relationships, the syntactic features of the hypothesis are critical. For example, *managerial trustworthy behavior* is the subject and *employees' trust in the manager* is the object of a verb phrase containing "is positively related to." Additionally, the moderation and mediation natures of the relationship constitute important behavioral knowledge. Based on several studies (e.g., Maynard et al. 2009; Tan et al. 2016), we posit that effectively extracting complex domain-specific relationships may require multi-stage approaches that combine statistics/ML- and linguistics-based methods. In this vein, SVM (Cortes and Vapnik 1995), motivated by statistical learning theory (Vapnik 1998), has been recognized as a strong performer for relation extraction (Zhou et al. 2005).

Synonymous Relationship Identification

The final task is to identify synonymous variables within an article or across articles. Linguistics- and statistics/ML-based methods are commonly fused to solve this problem, including lexical similarity analysis, latent semantic analysis, and semantic lexicon. Lexical similarity analysis assesses the degree of similarity between two texts at the lexical level and is a common ontology learning method (Kishore and Ramesh 2007; Gefen et al. 2020). Typical methods include minimum edit distance, which measures the minimum number of single-character edits (insertions, deletions, or substitutions) to convert one text string into another (Strube et al. 2002). Semantic lexicon is a popular resource for ontology learning (Wong et al. 2012) that consists of predefined concepts and relations and can be used for identifying terms, concepts, taxonomic, and non-taxonomic relations. Well-known semantic lexicons include WordNet (Fellbaum 1998; Miller 1995) and the Unified Medical Language System (Bodenreider 2004). To identify synonymous variables within an article, lexical similarity analysis and semantic lexicon could be used due to consistent naming conventions enforced in academic articles. For variables across articles, one state-of-the-art technique is the construct similarity algorithm proposed by Larsen and Bong (2016), which uses multiple semantic lexicons (e.g., WordNet) and a customized latent semantic analysis fused with other lexical similarity measures to extract the hypernym/hyponym relationships among constructs.

Testable Hypotheses

Testable hypotheses are intended to evaluate how well the proposed meta-design satisfies our meta-requirements (Walls et al. 1992). For the proposed design framework, this entails two aspects: the ability to extract behavioral knowledge from behavioral articles and the ability to enhance information-seeking outcomes. A multifaceted evaluation solution is needed to address these two aspects (Hevner et al. 2004; Gill and Hevner 2013). Specifically, the former hypothesis requires rigorous comparisons with alternative ontology learning techniques and systems, and the latter calls for user experiments to shed quantitative and qualitative light on when, to whom, how, and to what extent the proposed meta-design enhances behavioral researchers' information-seeking outcomes (Abbasi et al. 2018). However, all of the evaluation methods require instantiations of the proposed design framework as a basis to evaluate the effectiveness and applicability of our meta-design (Abbasi and Chen 2008).

The remainder of the paper is organized as follows. We first describe the TheoryOn systems (developed as an instantiation of the proposed BOLT design framework). The two ensuing sections provide experimental evaluations of the TheoryOn system and its underlying framework with regards to two testable hypotheses—the ability to extract behavioral knowledge and the ability to enhance information-seeking outcomes. The discussion of empirical insights and generalizability of the proposed design framework and instantiation are offered. We conclude with a summary of our research contributions and potential future directions.

THEORYON—AN INSTANTIATION OF THE PROPOSED DESIGN FRAMEWORK

Using the design guidelines prescribed by the BOLT framework, Figure 2 depicts the system diagram for the TheoryOn instantiation. Each article underwent *hypothesis extraction, variable extraction*, and *relationship extraction* using the techniques prescribed by our framework. Specifically, hypothesis extraction used a sentence classifier combining deep learning and rule-based extraction. For variable and relation extraction, we developed an SVM classifier with a deep learning-informed multi-stage tree composite kernel (DLMTCK). With the extracted variables and hypotheses, the theoretical network was assembled in the *theoretical network construction* step and was indexed and placed inside the TheoryOn *Search & Visualization* application. Both steps are instantiations for the *Synonymous Relationship Identification* task in BOLT. In the following, we discuss each step of TheoryOn in detail.

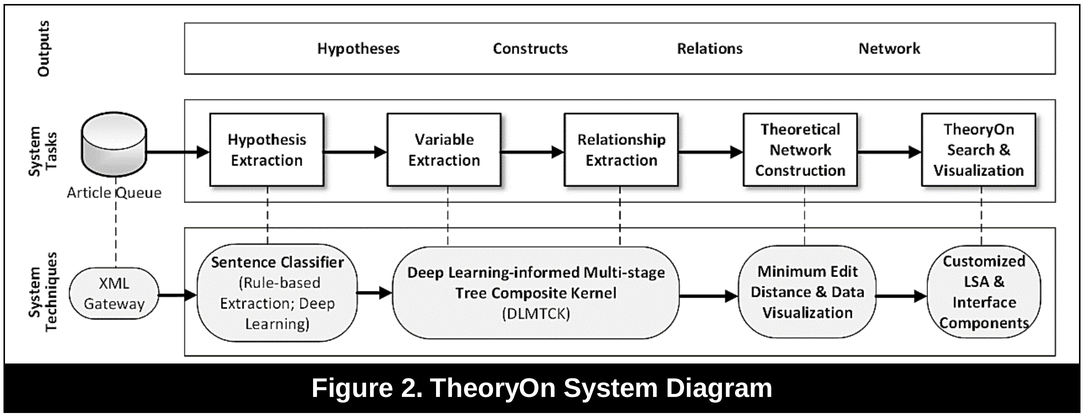

Outputs
Hypotheses
Constructs
Relations
Network
System Tasks
Article Queue
Hypothesis Extraction
Variable Extraction
Relationship Extraction
Theoretical Network Construction
TheoryOn Search & Visualization
System Techniques
XML Gateway
Sentence Classifier (Rule-based Extraction; Deep Learning)
Deep Learning-informed Multi-stage Tree Composite Kernel (DLMTCK)
Minimum Edit Distance & Data Visualization
Customized LSA & Interface Components

Figure 2. TheoryOn System Diagram

Hypothesis Extraction

Following the guidelines offered by the BOLT framework, TheoryOn utilizes a sentence classifier that couples rule-based and deep learning methods to extract hypotheses. We first used a rule-based approach to identify the hypothesis extraction rules. The details of the rule identification process are depicted in Appendix B. These extraction rules were then coupled with a deep learning method, depicted in Figure 3 and described below.

The word2vec method was used to create pre-trained word embeddings from article texts in order to efficiently incorporate distributional semantic information. This word embedding was then inputted into a Bi-LSTM to learn the long-distance dependencies among words. Additionally, we represented rule-based features as one-hot vectors, indicating whether the sentence contained an extraction pattern in Appendix B. We also incorporated the sentence order features, motivated by the fact that hypotheses often appear earlier in articles. All these features were concatenated and put into a dense layer to calculate the final classification probabilities. Later, in the evaluation section, we justify our design choices guided by BOLT by benchmarking our hybrid classifier against standard deep learning for text classification, a rule-based approach, and several feature-based methods.

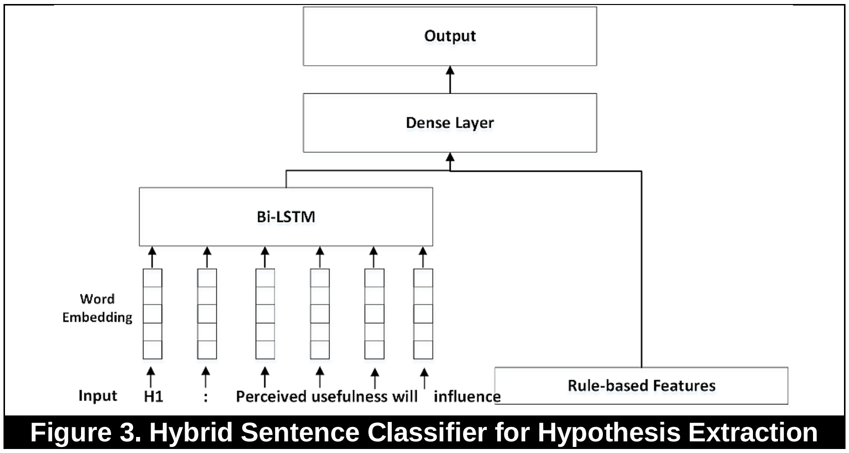
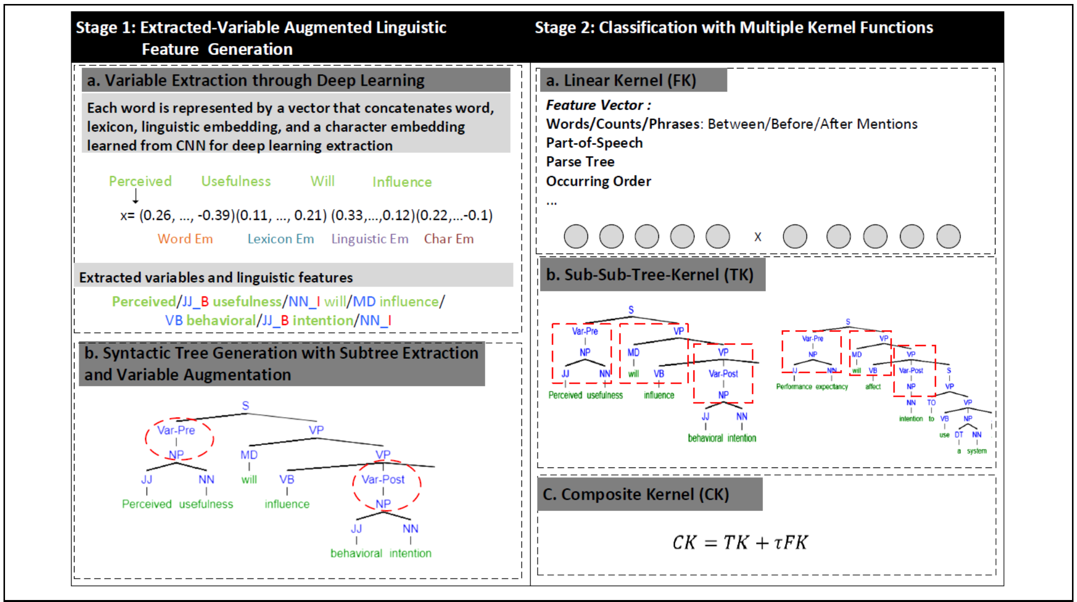


Figure 3. Hybrid Sentence Classifier for Hypothesis Extraction

Variable and Theoretical Relationship Extraction

According to the BOLT framework, the performance of theoretical relationship extraction was heavily dependent on the accuracy of variable extraction. Therefore, we combined these two steps to allow for fast iterations in model tuning. Accordingly, we propose a two-stage labeled tree kernel, with the first stage focusing on extracting rich linguistic patterns that are augmented by deep learning-informed variable extraction and the second stage encompassing an

SVM that fuses the domain-specific linguistic patterns in a composite kernel function. Our proposed DLMTCK approach is illustrated in Figure 4.


Stage 1: Extracted-Variable Augmented Linguistic Feature Generation

a. Variable Extraction through Deep Learning

Each word is represented by a vector that concatenates word, lexicon, linguistic embedding, and a character embedding learned from CNN for deep learning extraction

x= (0.26, ..., -0.39)(0.11, ..., 0.21) (0.33,...,0.12)(0.22,...-0.1)

Extracted variables and linguistic features

Perceived/JJ_B usefulness/NN_I will/MD influence/ VB behavioral/JJ_B intention/NN_I

b. Syntactic Tree Generation with Subtree Extraction and Variable Augmentation

Stage 2: Classification with Multiple Kernel Functions

a. Linear Kernel (FK)

Feature Vector :
Words/Counts/Phrases: Between/Before/After Mentions
Part-of-Speech
Parse Tree
Occurring Order
...

b. Sub-Sub-Tree-Kernel (TK)

C. Composite Kernel (CK)

$$CK = TK + \tau FK$$

Figure 4. A Deep Learning-Informed Multi-Stage Tree Composite Kernel (DLMTCK) Approach for Variable and Theoretical Relationship Extraction

Stage 1: Extracted-Variable Augmented Linguistic Feature Generation

The first step in Stage 1 involved extracting variables from the hypotheses using deep learning (Stage 1(a) of Figure 4). Guided by BOLT, the architecture included a character-level CNN, word embedding, lexicon embedding, and linguistic embedding. Character-level CNN modeled the morphological patterns for each word. After the CNN layer, each character was represented as a fixed dimensional vector. Max pooling was applied to aggregate the character-level embedding to the word level, which was then fed into the Bi-LSTM layer. Pre-trained word embedding was generated to represent the distributional semantics, which was learnable during the training phase. The semantic lexicon and linguistic embedding enriched domain adaptation and learned syntactic patterns specific to behavioral knowledge. Specifically, lexical embedding leveraged a one-hot vector to represent whether a word was contained in a behavioral lexicon. The behavioral lexicon was generated by deriving the top words (removing the stop words) from the training data. The linguistics feature embedding included POS and chunk tags and was initially represented as a one-hot vector, in which the size of the one-hot vector equaled the size of the POS tag and chunk set defined in the Penn Treebank II tag set.[6] The one-hot vectors for lexicon and linguistics features were fed into an embedding layer to generate a dense vector representation, which was concatenated with the pre-trained word embedding and fed into a Bi-LSTM layer, followed by a CRF layer to model the dependencies among IOB tags. The detailed deep learning architecture is presented in Figure 5.

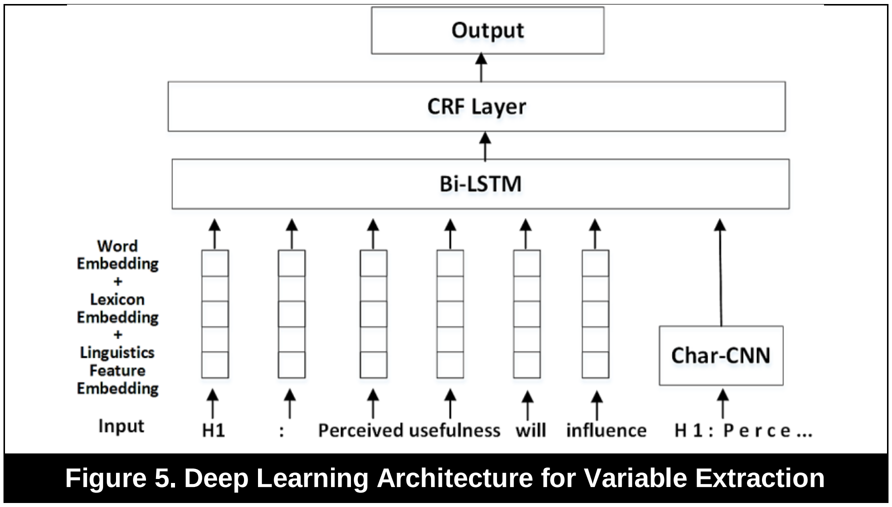


Figure 5. Deep Learning Architecture for Variable Extraction

Upon completion of variable extraction, in Stage 1b, we derived a syntactic tree for each hypothesis with subtree extraction and variable augmentation. Specifically, for any two variables in a hypothesis, we first extracted all subtrees that encompassed these two variables and then enriched the subtree with variable indicators. For example, in Figure 4, Stage 1(b), Var-Pre indicates the minimal phrase that contained the first variable. These rich subtree patterns were subsequently utilized in Stage 2, which is discussed later.

[6] https://www.clips.uantwerpen.be/pages/mbsp-tags

Stage 2: Classification With Multiple Kernel Functions

In the second stage, a composite kernel SVM was used to predict the theoretical relationships between variables. SVM uses the maximum margin principle to find two parallel hyperplanes that can divide a set of data points into two classes (e.g., having a particular relationship or not), in which the margin is defined by the perpendicular distance between these two parallel hyperplanes (Cristianini and Shawe-Taylor 2000). The hope is that the larger the margin, the smaller the generalization error. For our relation extraction problem, this is translated into finding the optimal hyperplanes for three relationship types: the main effect (variable1, variable2), moderation (variable1, moderator, variable2), and mediation (variable1, mediator, variable2). However, moderation and mediation involve three variables, whereas SVM relation extraction typically concerns the classification of relationships involving two variables. We therefore decomposed a ternary relationship for moderation into two binary relationships of moderation (moderator, variable) and one binary relationship of the main effect (variable1, variable2), and a ternary relationship for mediation into two binary relationships of mediation (mediator, variable) and one binary relationship of the main effect (variable1, variable2).

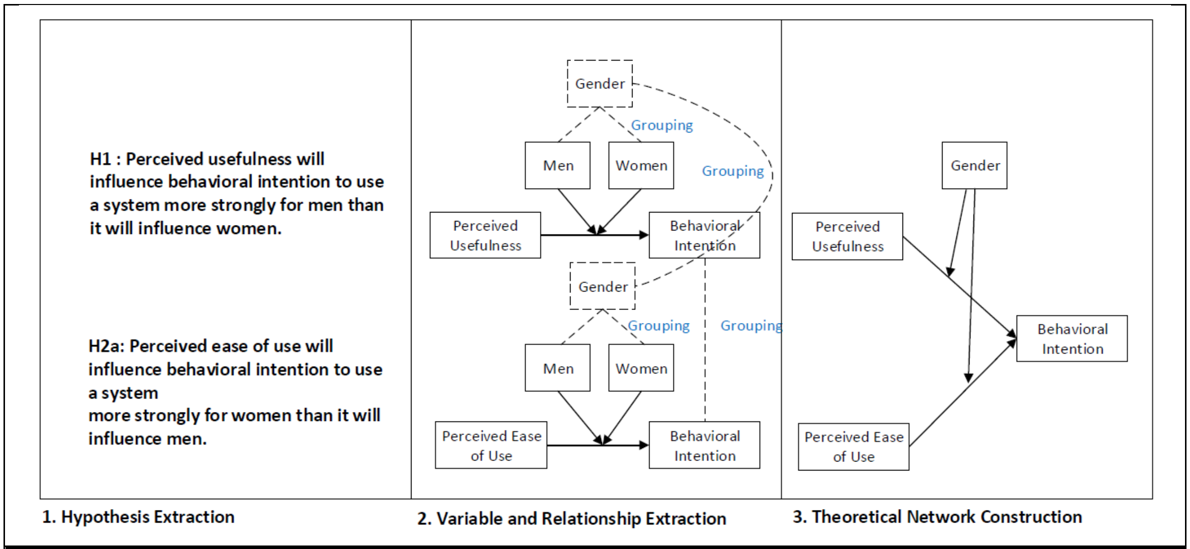
Consequently, for moderation and mediation relationships, we first classified the derived binary relationships and assembled them into ternary relationships. For each derived binary relationship, we used a composite kernel function to reflect the diverse linguistics patterns. In SVM, a kernel function measures the similarity between two feature vectors by mapping them to a higher-dimensional space so that an optimal hyperplane could be found. It can also be tailored to incorporate domain-specific knowledge (Burges 1998; Muller et al. 2001). Composite kernels are well suited to incorporate broad, relevant features while reducing the risk of over-fitting (Collins and Duffy 2002; Szafranski et al. 2010; Kitchens et al. 2018). Specifically, our composite kernel function (Stage 2c) is a linear combination of two kernel functions (Zhou et al. 2010), including a linear kernel function (Stage 2a) characterizing the flat linguistic patterns defined by Zhou et al. (2005) and a sub-tree kernel function (Stage 2b) incorporating the variable augmented subtree features from the first stage (Collins and Duffy 2002). A detailed description of these kernel functions is shown in Appendix C.

Once the relationship type of any variable pair in a hypothesis was determined, we consolidated the binary relationships into ternary relationships based on shared constructs. The three types can represent hypotheses with any number of variables. The unit of analysis for evaluation is thus based on how many of these types is extracted correctly.

Theoretical Network Construction

TheoryOn visualizes an article's theoretical network by grouping the variables shared across its hypotheses. For example, after variable and relation extraction, H1 and H2a in Venkatesh and Morris (2000) can be represented as solid boxes and arrows in Figure 6(2), in which men and women are the values of a gender variable moderating the relationship between perceived usefulness and behavioral intention, as well as that between perceived ease of use and behavioral intention.

In order to create a succinct theoretical network visualization, "men" and "women" are grouped as a "gender" variable through a semantic lexicon (e.g., women, men, boy, and girl are hyponyms for gender). Furthermore, two "behavioral intention" and two "gender" variables from H1 and H2a, respectively, are grouped together using lexical similarity analysis. Specifically, a minimum edit distance measure was used to calculate their similarity, with the threshold determined empirically using our training set. Combining all hypotheses through shared variables could automatically construct a theoretical network for each behavioral article.



H1 : Perceived usefulness will influence behavioral intention to use a system more strongly for men than it will influence women.

H2a: Perceived ease of use will influence behavioral intention to use a system more strongly for women than it will influence men.

Figure 6. Example of Theoretical Network Construction for Two Hypotheses in Venkatesh and Morris (2000)

TheoryOn Search and Visualization

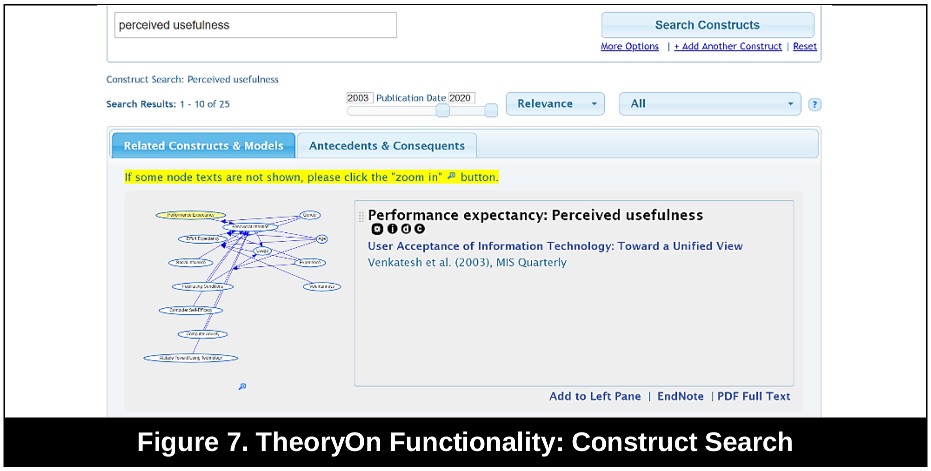
Automatically extracting theoretical networks allows users to conduct an ontology-centric search, in which a user types a variable as a search query to obtain a list of relevant theoretical networks. To accomplish this objective, synonymous relationships among variables from different articles should be identified. For example, the construct *performance expectancy* in Venkatesh et al. (2003) is synonymous with the construct *perceived usefulness* in Venkatesh and Morris (2000) despite comprising different words. When a user types the search query "perceived usefulness," both articles should be returned.

Following the guidance of BOLT, we utilized customized latent semantic analytics (LSA) (Larsen and Bong 2016) coupled with a standard Lucene keyword search algorithm (McCandless et al. 2010). TheoryOn users are given the option of selecting a keyword search or a combination of keyword and LSA search, serving multiple stages of the information-seeking process.

System Interface

Basing on common behavioral knowledge disembedding use cases related to the key output of the BOLT framework, we developed the following four system functionalities:

a) Construct Search. TheoryOn allows users to specify a construct in a search query in order to search articles that contain this construct or its synonymous constructs. Figure 7 shows a search for "perceived usefulness" using a combination of keyword and customized LSA search. Each retrieved construct is shown with the theoretical network it belongs to, with the target construct marked in yellow. For details, watch the video "TheoryOn: Synonymous Construct Search."[7]



perceived usefulness

Construct Search: Perceived usefulness

Search Results: 1 - 10 of 25

2003 Publication Date 2020

If some node texts are not shown, please click the "zoom in" button.

Performance expectancy: Perceived usefulness

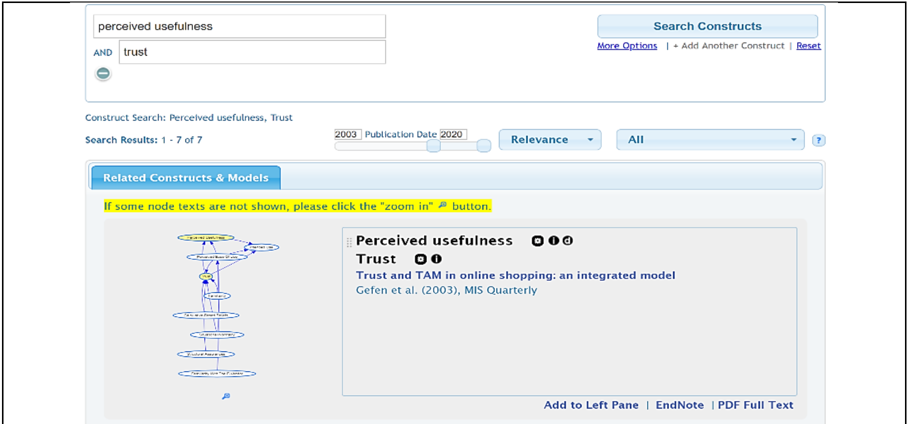
User Acceptance of Information Technology: Toward a Unified View

Venkatesh et al. (2003), MIS Quarterly

Figure 7. TheoryOn Functionality: Construct Search

b) Construct Pair Search. TheoryOn allows users to specify a construct pair in a search query and to find articles containing those two constructs (see Figure 8). The constructs (marked in yellow) and their relationships are shown in the extracted theoretical models in the left part of the search results. For more details, watch the video "TheoryOn: Construct-Pair Search."

[7] Video links lead to an anonymized YouTube channel, compliant with *MIS Quarterly*'s blind review policies. A prototype version of the tool is available at TheoryOn.org



perceived usefulness

AND trust

Construct Search: Perceived usefulness, Trust

Search Results: 1 - 7 of 7

2003 Publication Date 2020

If some node texts are not shown, please click the "zoom in" button.

Perceived usefulness

Trust

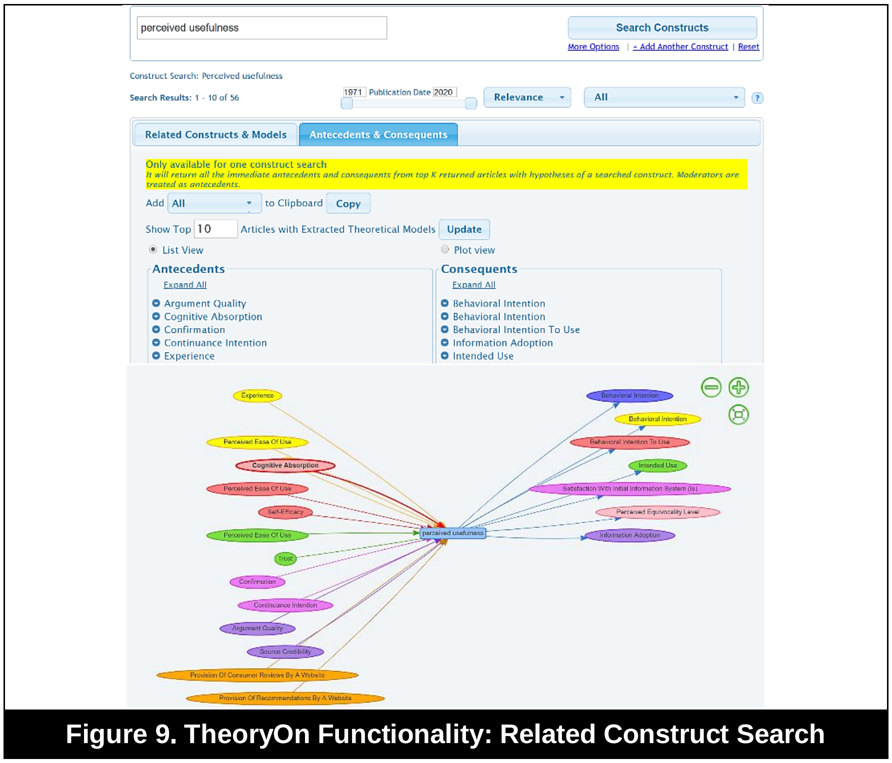
Trust and TAM in online shopping: an integrated model

Gefen et al. (2003), MIS Quarterly

Search Constructs
More Options | + Add Another Construct | Reset
Relevance
All
?
Related Constructs & Models
Add to Left Pane | EndNote | PDF Full Text

Figure 8. TheoryOn Functionality: Construct Pair Search

c) Theoretically Related Construct Search. This functionality allows inspection of the theoretical models containing a construct of interest (under "Antecedents" and "Consequents" section), as well as the examination of its antecedents and consequents in a list or plot view (Figure 9). TheoryOn takes the first *n* articles returned by the construct search and displays the antecedents to the construct searched for. It then does the same for the consequents. For more details, watch the video "TheoryOn: Theoretically Related Construct Search."



perceived usefulness

Construct Search: Perceived usefulness

Search Results: 1 - 10 of 56

1971 Publication Date 2020

Only available for one construct search

It will return all the immediate antecedents and consequents from top K returned articles with hypotheses of a searched construct. Moderators are treated as antecedents.

Add All to Clipboard Copy

Show Top 10 Articles with Extracted Theoretical Models Update

● List View ○ Plot view

Antecedents

Expand All

- Argument Quality
- Cognitive Absorption
- Confirmation
- Continuance Intention
- Experience

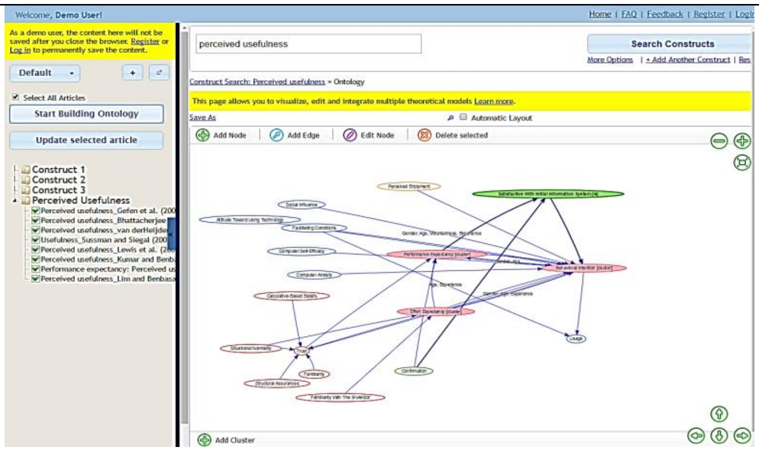
Consequents

Expand All

- Behavioral Intention
- Behavioral Intention
- Behavioral Intention To Use
- Information Adoption
- Intended Use

Figure 9. TheoryOn Functionality: Related Construct Search

d) Theory Integration. All the related theories can be saved in the left panel and visualized on the canvas (see Figure 10). A user can then integrate theories by clustering synonymous constructs, or the user can customize the theoretical networks by editing, deleting, or adding any nodes and links. For more details, watch the video "TheoryOn: Theory Integration."



Welcome, Demo User!

As a demo user, the content here will not be saved after you close the browser. Register or Log in to permanently save the content.

☑ Select All Articles

- Construct 1
- Construct 2
- Construct 3
- Perceived Usefulness
  - ☑ Perceived usefulness_Gefen et al. (200
  - ☑ Perceived usefulness_Bhattacherjee
  - ☑ Perceived usefulness_van derHeijde
  - ☑ Usefulness_Sussman and Siegal (200
  - ☑ Perceived usefulness_Lewis et al. (20
  - ☑ Perceived usefulness_Kumar and Benb
  - ☑ Performance expectancy: Perceived us
  - ☑ Perceived usefulness_Lim and Benbasa

perceived usefulness

Construct Search: Perceived usefulness » Ontology

This page allows you to visualize, edit and integrate multiple theoretical models Learn more.

Save As

□ Automatic Layout

Home | FAQ | Feedback | Register | Login
Default
+
Start Building Ontology
Update selected article
Search Constructs
More Options | + Add Another Construct | Res
Add Node
Add Edge
Edit Node
Delete selected
Perceived Enjoyment
Satisfaction With Initial Information System [a]
Social Influence
Attitude Toward Using Technology
Facilitating Conditions
Gender, Age, Voluntariness, Experience
Computer Self-Efficacy
Performance Expectancy [cluster]
Gender, Age
Behavioral Intention [cluster]
Computer Anxiety
Age, Experience
Gender, Age, Experience
Calculative-Based Beliefs
Effort Expectancy [cluster]
Usage
Situational Normality
Trust
Familiarity
Confirmation
Structural Assurances
Familiarity With The B-Vendor
Add Cluster

Figure 10. TheoryOn Functionality: Theory Integration

EVALUATION—EXPERIMENTS TO EXAMINE BEHAVIORAL KNOWLEDGE EXTRACTION PERFORMANCE

To demonstrate the effectiveness of the BOLT framework's design guidelines, as well as the proposed methods for hypothesis, variable, and relation extraction, we benchmarked our framework-guided techniques with alternative methods. For each task, the labeled set was divided into a training (60%), a development (20%), and a test set (20%). The training set was used to build the model, the development set was used to search for the optimal learning parameters, and the test set was used to report the model performance. The labeled data consisted of 69 articles from *MIS Quarterly*, 72 articles from *Information Systems Research*, and 145 articles from the *Journal of Applied Psychology* from 1980 to 2009. These three journals were chosen to illustrate the generalizability of our methods across multiple disciplines in behavioral research. Each article was labeled by two annotators with a combined 20+ years of research experience in behavioral research. The inter-rater reliability, measured by Cohen's kappa, for hypothesis, variable, and relationship extraction was 0.98, 0.75, and 0.82, respectively, indicating agreement levels that are substantial and close to being almost perfect (Landis and Koch 1977). Any disagreement between the two annotators was extensively discussed and resolved, resulting in 1,913 manually extracted hypotheses, 6,020 variable instances, and 3,135 basic relationships (binary or ternary). Two sets of experiments were performed: method comparison and system comparison. In the method comparison, we evaluated our hypotheses, variable, and relationship extraction methods against existing techniques. For the system comparison, TheoryOn's variable and relation extraction capabilities were compared with state-of-the-art text ontology learning systems. The details regarding the experiments are given next.

Method Comparison Experiments and Results

For the hypothesis extraction task, we compared the aforementioned hybrid sentence classifier with a rule-based method, feature-based classifiers, and deep learning classifiers. The feature-based classifiers included maximum entropy and naïve Bayes. The deep learning classifiers included a CNN classifier proposed by Kim (2014), an LSTM classifier by Tang et al. (2015), and a hybrid a BiLSTM-CNN classifier by Zhou et al. (2016). For variable and relation extraction, our proposed DLMTCK method was evaluated against state-of-the-art techniques. The variable extraction benchmarks included the BiLSTM + CRF classifier by Huang et al. (2015), the character-level CNN (CharCNN) + BiLSTM + CRF classifier by Ma and Hovy (2016), CRF, the domain relevance measure (DRM) (Jiang and Tan, 2010), the C/NC value (Drymonas et al. 2010), noun phrase term frequency–inverse document frequency (npTFIDF)

(Maedche and Staab 2000), and lexicon-driven concept identification through binary predicate (BPLex) (Oliveira et al. 2001). The classifier BiLSTM + CRF inputs a pre-trained word embedding to a BiLSTM layer and classifies IOB tags using a CRF decoder. The CharCNN + BiLSTM + CRF method extends this architecture by adding character-level embedding and a CNN to model the morphological patterns of each word. The DRM method identifies noun phrases and uses domain-specific lexicon coupled with a likelihood test measure to determine phrases that might constitute potential concepts. The C/NC value uses a heuristic measure to score candidate noun phrases based on co-occurrence patterns indicative of concept mentions. The npTFIDF extracts all noun phrases, removes articles and certain descriptive adjectives such as "several" and "many", and computes *tfidf* to eliminate irrelevant terms (i.e., those below a specified threshold). Finally, BPLex derives noun phrase patterns through a binary predicate function that usesuses part-of-speech tags and syntactic structures.

For relation extraction, the benchmarks included a linear SVM, the verb rule method (Jiang and Tan, 2010), association rule mining (Drymonas et al. 2010; Maedche and Staab 2000), and customized semantic template invoving lexico-syntactic patterns (LexSynPatt) (Oliveira et al. 2001; Vargas-Vera et al. 2001). The verb rule method utilized a predefined noun–verb–noun regular expressions that were capable of identifying non-taxonomic relations between constructs. Association rule mining was used to obtain noun-verb-noun rule sets encompassing antecedent constructs and consequent constructs with appropriate level of support and confidence levels. The LexSynPatt represented relation instances in the training set as item sets encompassing constructs, lexico-syntactic patterns such as verb-based binary predicates.

The experiment results are presented in Table 1. In terms of hypothesis extraction, the hybrid classifier performed better with precision, recall, and F_1-measure compared with the deep learning methods. The rule-based method had high precision and relatively low recall because of additional patterns residing in the test data. The feature-based methods, such as maximum entropy and naïve Bayes, performed worse than the deep learning methods did. In particular, naïve Bayes' recall is very low, mainly because of its sensitivity to contexts with a skewed prior class label distribution (such as our hypothesis extraction context).

Table 1. Method Comparison Results for Hypothesis, Variable, and Relationship Extraction

		Precision	Recall	F1
Hypothesis Extraction	*Hybrid*	**96.27%**	94.26%	**95.25%**
	BiLSTM-CNN	87.01%	92.69%	89.76%
	BiLSTM	94.04%	90.60%	92.29%
	CNN	80.26%	**95.56%**	87.25%
	Rule based	93.92%	88.77%	91.28%
	Maximum Entropy	95.12%	81.46%	87.76%
	Naïve Bayes	52.94%	14.15%	22.33%
Variable Extraction	*DLMTCK*	**77.07%**	**76.17%**	**76.61%**
	CNN + BiLSTM + CRF	74.89%	72.58%	73.72%
	BiLSTM + CRF	74.04%	72.50%	73.26%
	CRF	74.43%	70.58%	72.46%
	HMM	60.45%	55.42%	57.83%
	DRM	27.91%	48.92%	35.54%
	C/NC Value	24.44%	45.50%	31.80%
	npTFIDF	25.77%	45.75%	32.97%
	BPLex	23.81%	44.33%	30.98%
Relation Extraction	*DLMTCK*	**88.44%**	**80.98%**	**84.54%**
	Linear SVM	83.61%	78.04%	80.73%
	Verb Rule Method	63.24%	48.24%	54.73%
	Association Rules	65.33%	48.04%	55.37%
	LexSynPatt	72.29%	56.27%	63.29%

For variable extraction, the proposed DLMTCK method offered much better performance than the comparison methods did. The performances of CRF and deep learning methods are complementary in the sense that CRF could model class label dependency, whereas deep learning methods could effectively reprent input features. HMM was hampered by its reliance on feature token representations and its inability to consider long-distance interdependencies. Concept extraction methods from prior ontology learning studies are designed for extracting general-purpose concept, which may include valid concepts that are not behavioral constructs or noun phrases that are not exactly match the behavioral construct phrases. Nonetheless, these methods' over-reliance on general noun phrase extraction principles may not be suitable for behavioral ontology learning context.

For relation extraction, DLMTCK outperformed the linear SVM by about four percentage points on F-measure, demonstrating the value of the tree structure approach for the enhanced identification of construct relations. Once again, existing text ontology learning methods were designed for general-purpose relation extraction, which could include relationships that are outside of the theoretical contruct relationships or miss relationships that are not connected by verbs. Hence, they could not precisely and comprehensively represent the myriad relation patterns embodied in behavioral texts. Collectively, the results show the efficacy of the meta-design provided by the BOLT framework and demonstrate the utility of the proposed DLMTCK method for variable and relation extraction.

System Comparison Experiments and Results

To examine its system-level performance, TheoryOn was compared with existing text ontology learning systems. To select the most appropriate baseline systems, we used the evaluation guidelines proposed by Park et al. (2007), namely, general, extraction, and quality features, as inclusion criteria. Many systems we surveyed were not applicable because they lack sufficient extraction features such as extraction levels and degrees of automation (Park et al. 2007). For example, the FFCA system (Quan et al. 2004) and ASIUM (Faure and Poibeau 2000) do not tackle non-taxonomic relations, and DODDLE-OWL (Morita et al. 2006) uses a semi-automatic extraction process. Consequently, concept–relation–concept tuple-based ontology learning (CRCTOL; Jiang and Tan 2010), OntoGain (Drymonas et al. 2010), Text-To-Onto

(Maedche and Staab 2000), and TextStorm (Oliveira et al. 2001) were selected. Because these systems are not designed for behavioral ontology learning and may include other ontology extraction steps, we only selected their relevant components without modification for comparison. Specifically, CRCTOL automatically mines concepts and relations using DRM-based noun phrase extraction and predefined noun-verb-noun patterns. OntoGain uses a C/NC value-based noun phrase extraction algorithm coupled with an association rule mining-based relation extraction method. Text-To-Onto combines syntactic patterns of noun phrases with association rule mining, and TextStorm uses binary predicates for concept and relation extraction. Unlike the method evaluation, the system comparison examined the performance of the ontology learning pipelines, including the error/performance interaction effects between stages. As non-BOLT systems do not have formal hypothesis extraction mechanisms, we began with the extracted hypotheses and focused on the variable and relation extraction stages of the system pipelines.

The results are shown in Table 2. As expected, system pipeline performance was lower relative to the isolated testbed method results depicted in Table 1 because of error propagation. Consistent with the method experiments, TheoryOn offered better recall and F_1-measures for variable and relation extraction relative to the four comparison text ontology systems. The performance of generic text ontology systems confirms our initial belief that instantiations grounded in BOLT are necessary to make behavioral knowledge disembedding feasible. Next, we performed a user experiment and an applicability check to empirically demonstrate the practical downstream value of TheoryOn's hypothesis, variable, and relation extraction capabilities, which is discussed in the next section.

Table 2. System Comparison Results for Variable and Relationship Extraction

		Precision	Recall	F1
Variable Extraction	TheoryOn	**71.34%**	**70.33%**	**70.84%**
	CRCTOL	25.15%	43.17%	31.78%
	OntoGain	20.89%	39.83%	27.41%
	TextOnto	22.95%	39.92%	29.15%
	TextStorm	21.10%	38.42%	27.24%
Relation Extraction	TheoryOn	**74.05%**	**64.90%**	**69.17%**
	CRCTOL	38.44%	24.12%	29.64%
	OntoGain	34.85%	22.55%	27.38%
	TextOnto	35.15%	22.75%	27.62%
	TextStorm	31.54%	24.12%	27.33%

EVALUATION—USER EXPERIMENTS TO EXAMINE INFORMATION-SEEKING OUTCOMES

We conducted two user studies, namely a randomized user experiment and an applicability check, to evaluate TheoryOn's ability to enhance behavioral researchers' information-seeking outcomes quantitatively and qualitatively. Specifically, the randomized user experiment compares researchers' performance across four information-seeking tasks among TheoryOn, Google Scholar, and EBSCOhost. The applicability check uses the nominal group technique (NGT) to collect qualitative feedback from behavioral researchers in terms of when, to whom, and how TheoryOn might be beneficial to the information-seeking process.

Randomized User Experiment

We selected two full-text search engines, Google Scholar and the Business Source Complete database powered by EBSCOhost, as the benchmarking full-text search engines. Both of them represented, at the time of the experiment, the longest uninterrupted period of full-text coverage for *MIS Quarterly*, *Information Systems Research*, and the *Journal of Applied Psychology* (1990–2009). A total of 52 information systems and organizational behavior Ph.D. students from programs around the globe were randomly assigned to one of the three experimental groups (TheoryOn, EBSCOhost, or Google Scholar).

We designed the following four tasks for each participant to complete: *synonymous construct search*, *construct pair search*, *antecedent and consequent search*, and *theory integration*, each of which is a common scholarly information task for behavioral research. All four tasks were related to one theory, the technology acceptance model (TAM), to demonstrate a natural progression of knowledge acquisition, curation, and integration in an information-seeking process. TAM was selected because of high awareness, which set up a context in which users of Google Scholar and EBSCOhost were given every opportunity to perform at their peak.

The gold standard for each task was rigorously constructed by a team of two experienced faculty researchers, three doctoral students, and four senior research assistants (research assistants had at least 500 hours of experience in construct extraction from behavioral articles). Following Hevner et al. (2004) and Gill and Hevner (2013), we evaluated TheoryOn's performance using both objective and perceptual evaluations. The objective evaluation compared the construct, article, and theory retrieval performance, including precision and recall (Salton 1989), whereas the subjective evaluation examined the perceived utility of the artifact, reflected by the *perceived usefulness*, *perceived ease of use* and *behavioral intention* constructs from Davis (1989) and Venkatesh et al. (2003). The detailed experiment information regarding the randomization checks, task description, and evaluation procedure are depicted in Appendix D.

Construct and Theory Retrieval Performance

The results in Table 3 showed that the participants using TheoryOn attained F-measures that were 37% to 121% higher for all tasks, relative to the EBSCOhost and Google Scholar full-text search engines. Specifically, TheoryOn performed especially well in complex tasks, such as antecedent and consequent search, as well as in theory integration. These results demonstrate the viability of TheoryOn for potentially mitigating the knowledge inaccessibility problem that manifests during the scholarly information-seeking process. Compared with EBSCOhost and Google Scholar, TheoryOn could reduce false negatives in search results by up to 158%. This is because TheoryOn directly extracts hypotheses, constructs, and relationships and visualizes them in a user-friendly format, saving researchers precious time and effort otherwise expended extracting and processing behavioral knowledge from articles. The bandwidth freed up by TheoryOn's automated assistance allows users to process more articles (reducing false negatives) and shift their cognitive focus from labor-intensive manual extraction to better assessing the quality and relevance of the information examined (reducing the false positives). As a case in point, within the allotted four-hour time period, TheoryOn users were able to find, on average, 35.2 synonymous constructs, 35.2 antecedents, and 18 consequents, as well as integrate 13.8 theories—all nearly double compared with EBSCOhost and Google Scholar users. This finding is consistent with our prediction that IT artifacts that disembed behavioral knowledge from large-scale publications can allow users to focus on more value-added activities. The results of our subsequent qualitative applicability check further reinforced and underscored the speed, efficiency, and connection value proposition of TheoryOn during the processing stage of information-seeking behavior.

We also conducted an error analysis of TheoryOn users to understand the system bottleneck. On the one hand, failing to extract relevant constructs/relationships could result in false negatives in users' search results. In time-sensitive situations, examining all the results retrieved by TheoryOn may be challenging for users. This could explain why the user search recalls in Table 3 were lower than the method extraction recalls in Table 1. On the other hand, erroneous constructs/relationships undoubtedly led to some false positives in the search results. However, the users were able to assess and filter out many false positives via manual correction and refinement, which might explain why the users' search precision results were higher than the method extraction precision results presented earlier. Nonetheless, overall, the results of the user experiments suggest that TheoryOn has demonstrated its capabilities of lessening the cognitive load of manually processing knowledge and reducing false positives and false negatives in the scholarly information-seeking process. A future extension for this user experiment is allowing participants to combine and switch between whatever tools they may choose, such as Google Scholar, EBSCO, and TheoryOn, to yield valuable insights into the particular steps in the information-seeking process in which TheoryOn is most valuable.

Table 3. Percentage Retrieval Performance by Task

Task	TheoryOn (n = 18)			EBSCOhost (n = 17)			Google Scholar (n = 17)		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
1. Synonymous Construct Search	95.2	27.3	**40.1**	81.7	16.2	26.4	88.9	18.0	28.5
2. Construct Pair Search	76.7	43.9	**51.6**	72.0^{+}	24.7	34.9	51.2	34.7	37.2
3a. Antecedent Search	86.3	29.3	**41.5**	72.2	13.4	21.8	71.6	12.6	20.1
3b. Consequent Search	80.2	23.8	**34.7**	68.9	16.4	25.3	82.3^{+}	15.8	24.6
4. Theory Integration	77.4	25.4	**34.6**	61.9	16.0	23.9	62.9	9.8	15.6

Note: $^{+}$ not significantly different from TheoryOn ($p > 0.05$)

Perceived Utility

According to Table 4, across four tasks related to our proposed system functionalities, we found no significant difference in task experience (TE1–TE4; $p > 0.05$), but the perceived usefulness of TheoryOn for finding synonymous constructs, antecedents, and consequents and for extending theories was significantly better than that of EBSCOhost and Google Scholar, with a difference of 0.72 to 1.69 points on a seven-point Likert scale. Regarding overall utility perception at the system level, TheoryOn was considered to be significantly easier to use (EU) and useful (PU), whereas the behavioral intention to use the system (BI) was marginally significant. This marginal significance is likely due to TheoryOn not being publicly accessible at the time of the experiment; therefore, it was difficult for users to predict whether or not they would access the system in the next six months, which is the time frame used in the BI items.

Table 4. Perceived Usefulness Comparison of TheoryOn, EBSCOhost, and Google Scholar

	TheoryOn (n = 18)		EBSCOhost (n = 17)			Google Scholar (n = 17)		
Construct	Mean	SD	Mean	SD	Diff (t-stat)	Mean	SD	Diff (t_stat)
PU	5.92	0.73	5.01	1.01	3.04**	4.25	1.37	4.52***
EU	6.21	0.58	5.47	1.28	2.21*	5.78	1.27	2.14*
BI	5.57	1.21	4.84	1.26	1.74	6.57	2.24	−1.64
PU1	6.11	0.54	5.21	0.94	3.54**	5.37	1.42	2.07*
PU2	5.90	0.75	5.44	1.03	2.14*	4.82	1.46	2.77**
PU3	6.44	0.60	4.85	1.44	4.30***	4.96	1.45	4.01***
PU4	5.67	0.99	4.72	1.54	2.17**	4.89	1.57	2.42*
TE1	5.00	1.19	4.61	1.30	0.93	5.55	1.10	−1.41
TE2	5.69	0.92	5.24	1.14	1.29	5.06	1.29	1.66
TE3	5.26	1.35	4.82	1.24	0.99	4.61	1.53	1.34
TE4	4.22	1.46	4.63	1.47	−0.82	5.04	1.44	−1.67

Notes: 1. *p < 0.050; **p < 0.010; ***p < 0.001 2. PU: Perceived usefulness of the system; EU: ease of use of the system; BI: behavioral intention to use the system. PU1–4 are the perceived usefulness for each task. TE1–4 are the prior experience with each of the tasks; diff (t-stat) is the t statistics of EBSCOhost or Google Scholar compared with TheoryOn.

Applicability Check

We also conducted an applicability check to evaluate our system's *importance, accessibility*, and *suitability* to practitioners (Lukyanenko et al. 2019; Rosemann and Vessey 2008). We recruited 10 academic researchers at the assistant to full professor levels through an announcement to an academic listserv. The advertised inclusion criteria specified that they had to be social or behavioral researchers; had to hold a position equivalent to US titles of assistant, associate, or full professor; had to have published at least five academic articles, and had to be available for two 1.5-hour time slots.

Table 5. Summary of the Applicability Check for TheoryOn

Information-Seeking Behaviors	**Nominal Group Technique-Derived Information-Seeking Process**	**Supporting IT Artifacts**	**Quotes Related to TheoryOn**
Searching	• Formulate the problem/phenomenon • Identify the research questions • Identify the search terms • Search relevant articles • Screen for inclusion • Search articles related to the seed articles	Google Scholar Web of Science Medline Journal and association portals (AIS) ABI-Inform	"By identifying which papers are similar or redundant, it could save me a lot of time by quickly finding those new publications that I previously neglected." "TheoryOn could give doctoral students a decent start. It provides a quick and holistic view of a new area." "It could be a validation tool for reviewers to see whether a meta-analysis or literature review paper did a good job covering all the relevant papers." "It can work with citation management software, such as Mendeley, to accomplish a comprehensive solution to manage all related papers in a field." "If you start with a new research question, it is a very good tool to facilitate exploration and give a quick syncretization of the relevant research."
Accessing	• Access information systems or library portals	Web browser	
Processing	• Search the relevant keywords from selected articles • Annotate relevant arguments in articles • Discover contexts, variables, and theories • Extract citations • Synthesize arguments, variables, relations, theories, data, and findings • Categorize articles by usefulness and relevance • Build the discourse of the arguments and hypotheses		"TheoryOn really speeds up everything! It automatically extracts hypotheses, constructs, relationships, and models. So it facilitates synthesizing findings very well." "The most significant impact that TheoryOn has is six words: speeding up the evaluation of relevant papers. Traditional systems just present the abstracts. But you know, judging the relevance of a paper is more than its abstract. We need to look into variables, models, and findings, which TheoryOn has conveniently provided to us." "The system tremendously saves us time! This is very important. This morning, I was sitting in a panel. Someone talks about conducting a literature review of six hundred papers. The most challenging part is to code them. TheoryOn automatically extracts all the relevant pieces, so I can concentrate on the quality of the review rather than manually codifying the papers." "When I look at those models extracted by TheoryOn, I might start to think, hmm ... these relationships are missing. That triggers me to identify new research gaps." "TheoryOn can help highlight the key variables and constructs from the paper. It can also help me identify the most influential authors and papers — especially when I start a new domain." "TheoryOn's ability to pull all the papers and models together and extract all the relevant pieces is amazing!" "TheoryOn can help me link the constructs and save a lot of time. It just automatically does it!" "TheoryOn can help me build my own model. It can creatively suggest new papers or new models because it could find similar constructs between different papers." "If you already know the field, it helps you refine the research question, validate your understanding, and prioritize the most important papers."

The participants were engaged in two surveys, two one-hour NGT sessions, and a one-hour session and hands-on information search tasks for exposure to TheoryOn. The applicability check revealed 14 steps in the scholarly information-seeking process. For each step, the participants were asked to identify supporting IT artifacts. After being exposed to TheoryOn and completing the information retrieval tasks, the participants were asked to re-examine the information-seeking process and identify steps in which TheoryOn could be a significant help. The detailed process and materials are shown in Appendix E.

The NGT sessions were recorded, transcribed, and coded, and the results are summarized in Table 5. In general, the applicability check shed light on the scholarly information-seeking process and how it relates to the three information-seeking phases (searching, accessing, and processing), highlighted the potential value of construct-oriented search (and TheoryOn) during the processing phase, and touched on the potential for systems, such as TheoryOn, to complement existing options in the search phase.

Specifically, TheoryOn was considered *important* and *useful* for the scholarly information-seeking process, especially in the processing phase. The usefulness of TheoryOn is focused on saving time by immediately seeing the research models and being able to easily create new models through construct integration. Regarding accessibility, the participants applauded the user-friendly and intuitive interface: “wonderful to have a tool to visually support ontology construction” and “very interesting and useful—especially the graphic visualization.” Regarding suitability, the participants felt that TheoryOn could be especially useful and suitable for novice information seekers, especially those getting into a new field. Moreover, some participants felt that TheoryOn could help experienced researchers validate their understanding of a familiar field, refresh themselves on recent developments, and improve the overall quality of their scholarly pursuits. Some participants also noted that the tool could benefit reviewers by helping maintain quality while adding convenience in the peer-review process.

Additionally, they also commented on its complementarity to existing academic support IT artifacts. For example, they pointed out that “Google Scholar gave us coverage, but TheoryOn gave us precision,” and “TheoryOn has the potential to be implemented within the university library system.” Collectively, the applicability check validated the three phases of the information-seeking process, identified the stage in which TheoryOn could be especially helpful, and illustrated its importance, accessibility, and suitability.

DISCUSSION

In the following, we discuss the design science contribution of our paper by highlighting the accomplishments of the BOLT framework, TheoryOn instantiation, multifaceted evaluation, and generalizability of our proposed design artifacts. Finally, we discuss the potential impact of using the proposed design artifacts to mitigate the knowledge inaccessibility problem in behavioral research.

BOLT Framework. Following Walls et al. (1992), we proposed a BOLT design framework to offer concrete prescriptions for building artifacts capable of extracting specific ontology components related to behavioral knowledge disembedding. The method evaluation results demonstrated the superiority of the state-of-the-art prescriptions offered by the meta-design to support the nuances and complexities associated with the meta-requirements of BOLT. Furthermore, these results collectively underscored the feasibility of adopting the concept-centric perspective (Weber 2012) to disembed behavioral knowledge advocated by BOLT, where the extraction of hypotheses and constructs are critical precursors for disembedding behavioral knowledge.

TheoryOn System. The BOLT-guided TheoryOn system and its underlying extraction methods constitute important proof-of-concept artifacts. TheoryOn handily outperformed existing ontology learning systems and search engines. In particular, the randomized user experiment results showed that participants using TheoryOn attained F-measures that were 37% to 121% higher for all tasks, relative to the EBSCOhost and Google Scholar full-text search engines. Our applicability check shed light on the scholarly information-seeking process about when, to whom, and how construct-centric search engines might be beneficial, as well as the value proposition of tools such as TheoryOn. Overall, these results highlight the ability of BOLT-guided instantiation—TheoryOn—to extract behavioral knowledge from texts and to enhance information-seeking outcomes for behavioral researchers, verifying the importance of employing a multifaceted evaluation solution to demonstrate the practical value of TheoryOn.

Multifaceted Evaluation. Consistent with design principles (Hevner et al. 2004), we used a multifaceted evaluation to rigorously test each component of the proposed IT artifacts. The data mining experiments, randomized user experiment, and qualitative applicability check collectively offer additional empirical and qualitative insights that contribute to the academic literature on knowledge inaccessibility and information seeking in two ways:

1) Intelligent Text Analytics Can Alleviate Knowledge Inaccessibility. Our randomized user experiment showed that TheoryOn allowed its users to attain significantly better precision and recall, enabling behavioral researchers to access behavioral knowledge in an accurate and comprehensive manner. Prior work on the knowledge inaccessibility problem has largely focused on the comprehensiveness/recall problem, and our study confirmed the extent of this problem (Larsen and Bong 2016)—EBSCOhost and Google Scholar users were only able to retrieve between 9.8% and 34.7% percent of constructs on a fairly small article testbed (i.e., one favorable to higher recall rates). Interestingly, the user study also revealed lower precision rates. On three of the four tasks, the EBSCOhost and Google Scholar users were 6% to 49% lower on precision. This finding suggests that the bandwidth freed up by TheoryOn's automated assistance allows users to shift their cognitive focus from labor-intensive manual extraction to information quality and relevance examination, hence reducing false positives. Future design research on the knowledge inaccessibility problem should consider both precision and recall metrics as important considerations for artifact construction.

2) Empirical Evidence that BOLT Systems are Possible, Practical, and Valuable for Enhancing the Information-Seeking Process. The randomized user experiment and applicability check empirically revealed how the phases proposed by the information seeking literature (Meho and Tibbo 2003) are facilitated by the BOLT systems. Specifically, our randomized user experiment demonstrated that automatic behavioral knowledge extraction allows users to search for more articles (searching phase) and process more information in an accurate manner (processing phase). In addition, our qualitative applicability check validated the phases of the information seeing process and highlighted the potential value of complementing BOLT systems with existing artifacts to enhance the searching and processing phases. As far as we know, this article represents the first extensive examination of behavioral information-seeking processes and the potential for new, enabling design artifacts.

Generalizability. Our design artifacts could be applied to multiple behavioral and social disciplines such as behavioral medicine, psychology, education, and economics. They are also generalizable to NLP research (Abbasi and Chen 2008; Lau et al. 2012; Abbasi et al. 2019) as well as problem contexts and design solutions at the intersection of data, theory, and ML (Maass et al. 2018) in three ways:

1) Importance of Taking a Concept-Centric Perspective. The BOLT framework espouses the concept-centric perspective (Weber 2012), which showed that by focusing on effectively extracting hypotheses and constructs, the complex task of disembedding behavioral knowledge becomes viable. This simple and powerful idea of identifying key position statements and concepts nested within those statements can be generalized to many additional contexts such as philosophy and law, allowing for the development of robust IT artifacts for retrieving "locked" information and knowledge.

2) Deep Learning Methods for Complex NLP. The NLP research in IS has been dominated by topic categorization and sentiment polarity classification (Abbasi and Chen 2008; Lau et al. 2012; Abbasi et al. 2018; Zimbra et al. 2018). From an NLP perspective, these are relatively straightforward binary or multi-class classification problems (although accuracies for sentiment polarity detection remain challenging in certain domains). With the dramatic growth of a variety of user-generated text sources, methods capable of tackling more complex NLP problems such as knowledge extraction from behavioral data are at a premium (Ahmad et al. 2019). The results of our deep learning methods, fused with domain-specific features in a hybridized architecture, shed light on tackling complex NLP tasks in other fields such as biomedical text mining.

3) Holistic Evaluation for Design at the Intersection of Data, Theory, and Machine Learning. Evaluating design artifacts at the intersection of data, theory, and ML is particularly tricky (Prat et al. 2015; Maass et al. 2018). Our work is an example of such artifacts: the BOLT framework and TheoryOn instantiation rely on multiple behavioral and ontology learning theories, involve complex ML algorithms, and address structured and unstructured data throughout the design process. The empirical findings of our multifaceted evaluation solution revealed that a combination of data mining experiments, randomized user experiment, and qualitative applicability checks could help researchers reconcile competing approaches, identify design bottlenecks, and evaluate design solutions from diverse perspectives in this particular design context.

Impact of Mitigating Knowledge Inaccessibility. With the aid of our BOLT-guided TheoryOn search engine and combined with conventional search engines such as Google Scholar or EBSCOhost, the scholarly information-seeking process could be better supported, and the knowledge inaccessibility problem in behavioral research could be significantly mitigated. Specifically, with better awareness of existing constructs and relationships (as illustrated by high recalls in the user experiment), researchers are less likely to reinvent constructs or relationships already introduced by others, reducing wasted and redundant efforts as well as marginal research. Consequently, it would be easier to build a cumulative research tradition to ensure the persistent development and progression of a research discipline. Furthermore, by saving a lot of manual efforts of processing articles, researchers could improve the agility of the research topics and streamline their research process so as to quickly respond to environmental changes and grasp research opportunities. This research agility and efficiency could lead to profound monetary and societal benefits (e.g., speeding up behavioral intervention design for depression).

CONCLUSIONS AND FUTURE DIRECTIONS

Our contributions are threefold. First, we propose a BOLT design framework to guide the development of systems capable of behavioral knowledge disembedding and knowledge inaccessibility alleviation. Second, we instantiate our framework into a search engine artifact, TheoryOn, to show the applicability of the framework. TheoryOn also incorporates deep learning methods coupled with a composite kernel SVM to effectively extract hypotheses and constructs and their relations. Finally, through a series of data mining experiments, a randomized user experiment, and a qualitative applicability check, we offer additional empirical and qualitative insights that contribute to the academic literature on NLP research, design at the intersection of data, theory, and ML, information-seeking behaviors, and knowledge inaccessibility.

The level of success with which the hypothesis extraction, variable extraction, and relationship extraction were shown to work, and the improvements to which it led in a search experiment and applicability check, bodes well for the future. The solid performance of our design artifacts shows that future work is likely to be able to perform at such levels that behavioral knowledge disembedding will become the only option imaginable for evaluating past evidence. In fact, over the past 12 months, purely through word of mouth, the system has already garnered an impressive amount of usage. We believe these usage statistics would be further enhanced after a professional upgrade of the UI and UX interface (Kumar et al. 2004).

- *Engagement* – Over 4,000 engaged users who performed an average of 11 major actions per session, with an average session duration of nearly 5 ½ minutes, and who in total ran over 17,500 unique construct searches.
- *Reach* – These engaged users came from 459 academic institutions across 125 countries, with over 75% of users coming from Europe and Asia.

In this era of profound digital transformation, automation is disrupting various manual processes. Our proposed BOLT framework could have the potential to enable much more accurate literature search, automatic literature review, and automatic meta-analysis, as well as enable us to chart future directions for these disciplines more efficiently. We expect to work with experts in the biological and computer sciences to further refine and improve the framework proposed here and believe that the IS discipline is the natural home for this kind of work because of our understanding of design science, behavioral approaches, and NLP.

ACKNOWLEDGEMENTS

We thank the U.S. National Science Foundation for partial research support under the following grants: SBE-0965338, IIS-1816504, CCF-1629450, BDS-1636933, IIS-1553109, and IIS-1236970. We thank the University of Colorado – the work was also partially supported by the Center for Business Analytics at the University of Virginia.

REFERENCES

Abbasi, A., and Chen, H. 2008. “CyberGate: A Design Framework and System for Text Analysis of Computer-Mediated Communication,” *MIS Quarterly* (32:4), pp. 811-837.

Abbasi, A., Zhou, Y., Deng, S., & Zhang, P. 2018. “Text Analytics to Support Sense-Making in Social Media: a Language-Action Perspective,” *MIS Quarterly* (42:2), pp. 427-464.

Abbasi, A., Li, J., Adjeroh, D., Abate, M., and Zheng W. 2019 “Don’t Mention It? Analyzing User-generated Content Signals for Early Adverse Event Warnings,” *Information Systems Research*, (30:3), pp. 1007-1028.

Ahmad, F., Abbasi, A., Li, J., Dobolyi, D., Netemeyer, R., Clifford, G., and Chen, H. 2020. “A Deep Learning Architecture for Psychometric Natural Language Processing,” *ACM Transactions on Information Systems*, (38:1), article no. 6.

Arnulf, J. K., Larsen, K. R., Martinsen, Ø. L., and Bong, C. H. 2014. "Predicting Survey Responses: How and Why Semantics Shape Survey Statistics on Organizational Behaviour," *PloS One* (9:9), p. e106361.

Arnulf, J. K., Larsen, K. R., Martinsen, Ø. L., and Egeland, T. 2018. "The Failing Measurement of Attitudes: How Semantic Determinants of Individual Survey Responses Come to Replace Measures of Attitude Strength," *Behavior Research Methods*, pp. 1-21.

Ajzen, I. 1991. “The Theory of Planned Behavior,” *Organizational Behavior and Human Decision Processes* (50:2), pp. 179-211.

Asim, M. N., Wasim, M., Khan, M. U. G., Mahmood, W., and Abbasi, H. M. 2018. “A Survey of Ontology Learning Techniques and Applications,” *Database* (2018:1), p. 101.

Baron, R. M., and Kenny, D. A. 1986. “The Moderator–Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations,” *Journal of Personality and Social Psychology* (51:6), pp. 1173-1182.

Beel, J., and Gipp, B. 2010. “On the Robustness of Google Scholar against Spam,” in *Proceedings of the 21st ACM Conference on Hypertext and Hypermedia*, New York, NY: ACM, pp. 297-298.

Berger, A. L., Pietra, V. J. D., and Della, P. S. A. 1996. “A Maximum Entropy Approach to Natural Language Processing,” *Computational Linguistics* (22:1), pp. 39-71.

Biemann, C. 2005. “Ontology Learning from Text: A Survey of Methods,” *LDV Forum*, pp. 75-93.

Bodenreider, O. 2004. “The Unified Medical Language System (UMLS): Integrating Biomedical Terminology,” *Nucleic Acids Research* (32:suppl_1), pp. D267-D270.

Boeker, M., Vach, W., and Motschall, E. 2013. “Google Scholar as Replacement for Systematic Literature Searches: Good Relative Recall and Precision Are Not Enough,” *BMC Medical Research Methodology* (13:1), p. 131.

Buitelaar, P., Cimiano, P., and Magnini, B. (eds.) 2005. *Ontology Learning from Text: Methods, Evaluation and Applications*. Amsterdam, The Netherlands: IOS Press.

Bunge, M. 1977. “Emergence and the Mind,” *Neuroscience* (2:4), pp. 501-509.

Bunge, M. A. 1979. *Treatise on Basic Philosophy: Ontology II: A World of Systems*. Dordrecht, Holland: D. Reidel Publishing Company.

Burges, C. J. 1998. “A Tutorial on Support Vector Machines for Pattern Recognition,” *Data Mining and Knowledge Discovery* (2:2), pp. 121-167.

Bushman, B. J., and Wells, G. L. 2001. “Narrative Impressions of Literature: The Availability Bias and the Corrective Properties of Meta-Analytic Approaches,” *Personality and Social Psychology Bulletin* (27:9), pp. 1123-1130.

Cho, K., Van Merriënboer, B., Bahdanau, D., and Bengio, Y. 2014. “On the Properties of Neural Machine Translation: Encoder-Decoder Approaches,” *arXiv:1409.1259*.

Collins, M., and Duffy, N. 2002. “New Ranking Algorithms for Parsing and Tagging: Kernels over Discrete Structures, and the Voted Perceptron,” in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Stroudsburg, PA: Association for Computational Linguistics, pp. 263-270.

Compeau, D. R., and Higgins, C. A. 1995. “Computer Self-Efficacy: Development of a Measure and Initial Test,” *MIS Quarterly* (19:2), pp. 189-211.

Corley, K. G., and Gioia, D. A. 2011. “Building Theory about Theory Building: What Constitutes a Theoretical Contribution? ” *Academy of Management Review*, (36:1), pp. 12-32.

Cortes, C., and Vapnik, V. 1995. “Support-Vector Networks,” Machine Learning (20:3), pp. 273-297.

Cristianini, N., and Shawe-Taylor, J. 2000. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge University Press.

Davis, F. D. 1989. “Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology,” *MIS Quarterly* (13:3), pp. 319-340.

Davis, F. D., Bagozzi, R. P., and Warshaw, P. R. 1989. “User Acceptance of Computer Technology: A Comparison of Two Theoretical Models,” *Management Science* (35:8), pp. 982-1003.

DeLone, W. H., and McLean, E. R. 1992. “Information Systems Success: The Quest for the Dependent Variable,” *Information Systems Research* (3:1) pp. 60-95.

Doll, W. J., and Torkzadeh, G. 1988. “The Measurement of End-User Computer Satisfaction,” *MIS Quarterly* (12:2), pp. 259-274.

Drymonas, E., Zervanou, K., and Petrakis, E. G. 2010. “Unsupervised Ontology Acquisition from Plain Texts: The Ontogain System,” in *NLDB ’10 Proceedings of the Natural Language Processing and Information Systems, and 15th International Conference on Applications of Natural Language to Information Systems*, Berlin: Springer-Verlag, pp. 277-287.

Ellis, D. 1989. “A Behavioural Approach to Information Retrieval System Design,” *Journal of Documentation* (45:3), pp. 171-212.

Eyre, T. A., Ducluzeau, F., Sneddon, T. P., Povey, S., Bruford, E. A., and Lush, M. J. 2006. “The Hugo Gene Nomenclature Database, 2006 Updates,” *Nucleic Acids Research* (34:suppl_1), pp. D319-D321.

Faure, D., and Poibeau, T. 2000. “First Experiments of Using Semantic Knowledge Learned by ASIUM for Information Extraction Task Using Intex,” in *Proceedings of the First International Conference on Ontology Learning ECAI-2000 Workshop*, Aachen, Germany: CEUR-WS.org, pp. 7-12.

Fellbaum, C. 1998. “A Semantic Network of English Verbs,” in *WordNet: An Electronic Lexical Database*, C. Fellbaum (ed.), Cambridge, MA: MIT Press, pp. 153-178.

Gefen, D., Karahanna, E., and Straub, D. W. 2003. “Trust and TAM in Online Shopping: An Integrated Model,” *MIS Quarterly* (27:1) pp. 51-90.

Gefen, D., Endicott, J. E., Fresneda, J. E., Miller, J. L., and Larsen, K. R. "A Guide to Text Analysis with Latent Semantic Analysis in R with Annotated Code: Studying Online Reviews and the Stack Exchange Community," 2020.

Gill, T. G. 2001. “What’s an MIS Paper Worth? an Exploratory Analysis,” *ACM SIGMIS Database* (32:2), pp. 14-33.

Gill, T. G., and Hevner, A. R. 2013. “A Fitness-Utility Model for Design Science Research,” *ACM Transactions on Management Information Systems (TMIS)* (4:2), pp. 5:1-5:24.

Goodhue, D. L., and Thompson, R. L. 1995. “Task-Technology Fit and Individual Performance,” *MIS Quarterly* (19:2), pp. 213-236.

Gregor, S. 2006. “The Nature of Theory in Information Systems,” *MIS Quarterly* (30:3), pp. 611-642.

Gregor, S., and Hevner, A. R. 2013. “Positioning and Presenting Design Science Research for Maximum Impact,” *MIS Quarterly* (37:2), pp. 337-355.

Hearst, M. 1998. “Automated Discovery of WordNet Relations,” in *WordNet: An Electronic Lexical Database and Some of its Applications*, C. Fellbaum (ed.) Cambridge, MA: MIT Press, pp. 131-152.

Hevner, A., March, S., Park, J., and Ram, S. 2004. “Design Science in Information Systems Research,” *MIS Quarterly* (28:1), pp. 75-105.

Hobbs, J., and Riloff, E. 2010. “Information Extraction,” in *Handbook of Natural Language Processing,* N. Indurkhya and F. J. Damerau (eds.). Boca Raton, FL: CRC Press, pp. 511-532.

Hochreiter, S., and Schmidhuber, J. 1997. “Long Short-Term Memory,” *Neural Computation* (9:8), pp. 1735-1780.

Huang, Z., Xu, W., and Yu, K. 2015. “Bidirectional LSTM-CRF Models for Sequence Tagging,” *arXiv:1508.01991*.

Iacovou, C. L., Benbasat, I., and Dexter, A. S. 1995. “Electronic Data Interchange and Small Organizations: Adoption and Impact of Technology,” *MIS Quarterly* (19:4) pp. 465-485.

Im, G., and Straub, D. 2012. “Building Cumulative Tradition in Organization Science: A Methodology for Utilizing External Validity for Theoretical Generalization.” GSU, p. 36.

Jiang, X., and Tan, A. H. 2010. “CRCTOL: A Semantic-Based Domain Ontology Learning System,” *Journal of the Association for Information Science and Technology* (61:1), pp. 150-168.

Keen, P. 1980. “MIS Research: Reference Disciplines and a Cumulative Tradition,” in *Proceedings of the 1st International Conference on Information Systems (ICIS)*, Copenhagen, Denmark, October 18–20, 2006, pp. 9-18.

Kim, Y. 2014. “Convolutional Neural Networks for Sentence Classification,” *arXiv:1408.5882*.

Kitchens, B., Dobolyi, D., Li, J., and Abbasi, A. 2018. “Advanced Customer Analytics: Strategic Value through Integration of Relationship-Oriented Big Data,” *Journal of Management Information Systems*, 35(2), pp. 540-574.

Krosgaard, M. A., Brodt, S. E., and Whitener, E. M. 2002. “Trust in the Face of Conflict: The Role of Managerial Trustworthy Behavior and Organizational Context,” *Journal of Applied Psychology* (87:2), pp. 312-319.

Kumar, R. L., Smith, M. A., and Bannerjee, S. 2004. “User interface features influencing overall ease of use and personalization,” *Information & Management* (41:3), pp. 289-302.

Lafferty, J., McCallum, A., and Pereira, F. 2001. “Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data,” in *Proceedings of 18th International Conference on Machine Learning*, C. E. Brodley and A. P. Danyluk (eds.), San Francisco, CA: Morgan Kaufmann Publishers, pp. 282-289.

Landis, J. R., and Koch, G. G. 1977. “The Measurement of Observer Agreement for Categorical Data, ” *Biometrics* (33:1), pp. 159-174.

Larsen, K. R., and Bong, C. H. 2016. “A Tool for Addressing Construct Identity in Literature Reviews and Meta-Analyses,” *MIS Quarterly* (40:3), pp. 529-551.

Larsen, K. R., Hovorka, D. S., West, J. D., and Dennis, A. R. 2019. “Understanding the Elephant: A Discourse Approach to Corpus Identification for Theory Review Articles,” Journal of the Association for Information Systems, in press.

Larsen, K. R., Voronovich, Z. A., Cook, P. F., and Pedro, L. W. 2013. "Addicted to Constructs: Science in Reverse?," *Addiction* (108:9), pp. 1532-1533.

Lau, R. Y., Liao, S. S., Wong, K.-F., and Chiu, D. K. 2012. “Web 2.0 Environmental Scanning and Adaptive Decision Support for Business Mergers and Acquisitions,” *MIS Quarterly* (36:4), pp. 1239-1268.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. 1998. “Gradient-Based Learning Applied to Document Recognition,” *Proceedings of the IEEE* (86:11), pp. 2278-2324.

Luo, Y., Uzuner, Ö., and Szolovits, P. 2016. "Bridging Semantics and Syntax with Graph Algorithms—State-of-the-Art of Extracting Biomedical Relations," *Briefings in Bioinformatics* (18:1), pp. 160-178.

Lukyanenko, R. Parsons, J., Wiersma, Y. and Maddah, M. 2019 "Expecting the Unexpected: Effects of Data Collection Design Choices on the Quality of Crowdsourced User-Generated Content," *MIS Quarterly*, forthcoming

Ma, X., and Hovy, E. 2016. "End-to-End Sequence Labeling Via Bi-Directional LSTM-CNNs-CRF," *arXiv:1603.01354*.

Maass, W., Parsons, J., Purao, S., Storey, V. C., and Woo, C. 2018. "Data-Driven Meets Theory-Driven Research in the Era of Big Data: Opportunities and Challenges for Information Systems Research," *Journal of the Association for Information Systems* (19:12), pp. 1253-1273.

Maedche, A., and Staab, S. 2000. "Mining Ontologies from Text, " In *International Conference on Knowledge Engineering and Knowledge Management*, Berlin, Heidelberg: Springer, pp.189-202.

March, S. T., and Smith, G. 1995. "Design and Natural Science Research on Information Technology," *Decision Support Systems* (15:4), pp. 251-266.

Maynard, D., Funk, A., and Peters, W. 2009. "Using Lexico-Syntactic Ontology Design Patterns for Ontology Creation and Population," in *Proceedings of the 2009 International Conference on Ontology Patterns-Volume 516*, Aachen, Germany: CEUR-WS.org, pp. 39-52.

McCandless, M., Hatcher, E., and Gospodnetic, O. 2010. *Lucene in Action: Covers Apache Lucene 3.0*, Stamford, CT: Manning Publications Co.

McKnight, D. H., Choudhury, V., and Kacmar, C. 2002. "Developing and Validating Trust Measures for E-Commerce: An Integrative Typology," *Information Systems Research* (13:3), pp. 334-359.

Meho, L. I., and Tibbo, H. R. 2003. "Modeling the Information-Seeking Behavior of Social Scientists: Ellis's Study Revisited," *Journal of the American Society for Information Science and Technology* (54:6), pp. 570-587.

Mikolov, T., Karafiát, M., Burget, L., Černocký, J., and Khudanpur, S. 2010. "Recurrent Neural Network Based Language Model," in *11th Annual Conference of the International Speech Communication Association*, Baixas, France: International Speech Communications Association, pp. 1045-1048.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. 2013. "Distributed Representations of Words and Phrases and Their Compositionality," in *NIPS '13 Proceedings of the 26th International Conference on Neural Information Processing Systems*, Lake Tahoe, NV: Curran Associates, Inc., pp. 3111-3119.

Miller, G. A. 1995. "Wordnet: A Lexical Database for English," *Communications of the ACM* (38:11), pp. 39-41.

Missikoff, M., Navigli, R., and Velardi, P. 2002. "Integrated Approach to Web Ontology Learning and Engineering," *Computer* (35:11), pp. 60-63.

Morita, T., Fukuta, N., Izumi, N., and Yamaguchi, T. 2006. "DODDLE-OWL: A Domain Ontology Construction Tool with Owl," in *ASWC '06 Proceedings of the First Asian Conference on the Semantic Web*, Berlin: Springer-Verlag, pp. 537-551.

Muller, K.-R., Mika, S., Ratsch, G., and Scholkopf, B. 2001. "An Introduction to Kernel-Based Learning Algorithms," *IEEE Transactions on Neural Networks* (12:2), pp. 181-201.

Nelson, R. R. 1991. "Educational Needs as Perceived by IS and End-User Personnel: A Survey of Knowledge and Skill Requirements," *MIS Quarterly* (15:4), pp. 503-525.

Ng, A., and Jordan, M. 2002. "On Discriminative vs. Generative Classifiers: A Comparison of Logistic Regression and Naïve Bayes," in *NIPS '01 Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic*, Cambridge, MA: MIT Press, pp.841-848

Nickerson, R. S. 1998. "Confirmation Bias: A Ubiquitous Phenomenon in Many Guises," *Review of General Psychology* (2:2), pp. 175-220.

Oliveira, A., Pereira, F. C., and Cardoso, A. 2001. "Automatic Reading and Learning from Text," in *Proceedings of the International Symposium on Artificial Intelligence (ISAI)*, Kolhapur, India, December 2001.

Park, J., Cho, W., and Rho, S. 2007. "Evaluation Framework for Automatic Ontology Extraction Tools: An Experiment, " *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems*: Springer, pp. 511-521.

Parsons, J. and Wand, Y. (2013). "Extending Principles of Classification from Information Modeling to Other Disciplines," *Journal of the Association for Information Systems*, 14(4), pp. 245-273.

Peffers, K. 2002. "Perishable Research and the Need for a New Kind of IS Journal," *JITTA: Journal of Information Technology Theory and Application* (4:1), p. V.

Pennington, J., Socher, R., and Manning, C. 2014. "Glove: Global Vectors for Word Representation," *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Stroudsburg, PA: Association for Computational Linguistics, pp. 1532-1543.

Popper, K. 1959. *The Logic of Scientific Discovery*. New York: Basic Books.

Prat, N., Comyn-Wattiau, I., and Akoka, J. 2015. "A Taxonomy of Evaluation Methods for Information Systems Artifacts," *Journal of Management Information Systems* (32:3), pp. 229-267.

Quan, T. T., Hui, S. C., Fong, A. C. M., and Cao, T. H. 2004. "Automatic Generation of Ontology for Scholarly Semantic Web," *International Semantic Web Conference*, Berlin: Springer-Verlag, pp. 726-740.

Quirchmayer, G., Basl, J., You, I., Xu, L., and Weippl, E. 2012. *Multidisciplinary Research and Practice for Information Systems: International Cross Domain Conference and Workshop on Availability, Reliability, and Security*, Prague, Czech Republic, August 20-24, 2012.

Rabiner, L. 1989. "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," in *Proceedings of the IEEE* (77:2), pp. 257-286.

Rai, A. 2017. "Editor's Comments: Avoiding Type III Errors: Formulating IS Research Problems That Matter," *MIS Quarterly* (41:2), pp. iii-vii.

Rosemann, M., and Vessey, I. 2008. "Toward Improving the Relevance of Information Systems Research to Practice: The Role of Applicability Checks," *MIS Quarterly* (3:1), pp. 1-22.

Salton, G. 1989. *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*," Boston, MA: Addison-Wesley Longman Publishing Co., Inc.

Schryen, G., Benlian, A., Rowe, F., Shirley, G., Larsen, K., Petter, S., Paré, G., Wagner, G., Haag, S., and Yasasin, E. 2017. "Literature Reviews in IS Research: What Can Be Learnt from the Past and Other Fields?," *Communications of the Association for Information Systems* (41:1), p. 30.

Sharman, R., Kishore, R., and Ramesh, R. (Eds.) 2007. *Ontologies: A Handbook of Principles, Concepts and Applications in Information Systems*. New York: Springer Science & Business Media.

Soper, D. S., and Turel, O. 2015. "Identifying Theories Used in North American IS Research: A Bottom-up Computational Approach," in *48th Hawaii International Conference on System Sciences*. Washington, DC: IEEE, pp. 4948-4958.

Spell, C. S. 2001. "Management Fashions – Where Do They Come From, and Are They Old Wine in New Bottles?" *Journal of Management Inquiry* (10:4), pp. 348-373.

Strube, M., Rapp, S., and Müller, C. 2002. "The Influence of Minimum Edit Distance on Reference Resolution," in *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing-Volume 10*, Stroudsburg, PA: Association for Computational Linguistics, pp. 312-319.

Szafranski, M., Grandvalet, Y., and Rakotomamonjy, A. 2010. "Composite Kernel Learning," *Machine Learning* (79:1-2), pp. 73-103.

Tan, S. S., Lim, T. Y., Soon, L.-K., and Tang, E. K. 2016. "Learning to Extract Domain-Specific Relations from Complex Sentences," *Expert Systems with Applications* (60), pp. 107-117.

Tang, D., Qin, B., Feng, X., and Liu, T. 2015. "Effective LSTMS for Target-Dependent Sentiment Classification," *arXiv:1512.01100.*

Trinh-Phuong, T., Molla, A., and Peszynski, K. 2012. "Enterprise Systems and Organizational Agility: A Review of the Literature and Conceptual Framework," *Communications of the Association for Information Systems* (31:1), pp. 167-193.

Vargas-Vera, M., Domingue, J., Kalfoglou, Y., Motta, E., and Buckingham Shum, S. 2001. "Template-Driven Information Extraction for Populating Ontologies," In *IJCAI'01 Workshop on Ontology Learning*, Seattle, WA.

Vapnik, V. N. 1998. *Statistical Learning Theory*. New York: John Wiley & Sons.

Venkatesh, V., and Morris, M. G. 2000. "Why Don't Men Ever Stop to Ask for Directions? Gender, Social Influence, and Their Role in Technology Acceptance and Usage Behavior," *MIS Quarterly* (24:1), pp. 115-139.

Venkatesh, V., Morris, M. G., Davis, G. B., and Davis, F. D. 2003. "User Acceptance of Information Technology: Toward a Unified View," *MIS Quarterly* (27:3), pp. 425-478.

Walls, J. G., Widmeyer, G. R., and El Sawy, O. A. 1992. "Building an Information System Design Theory for Vigilant EIS," *Information Systems Research* (3:1), pp. 36-59.

Weber, R. 2012. "Evaluating and Developing Theories in the Information Systems Discipline," *Journal of the Association for Information Systems* (13:1), pp. 1-30.

Webster, J., and Watson, R. T. 2002. "Analyzing the Past to Prepare for the Future: Writing a Literature Review," *MIS Quarterly* (26:2), pp. XIII-XXIII.

White, R. 2013. "Beliefs and Biases in Web Search," in *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, New York: ACM, pp. 3-12.

Wong, W., Liu, W., and Bennamoun, M. 2012. "Ontology Learning from Text: A Look Back and Into the Future," *ACM Computing Surveys (CSUR)* (44:4), p. 20.

Yadav, V., and Bethard, S. 2018. "A Survey on Recent Advances in Named Entity Recognition from Deep Learning Models," in *Proceedings of the 27th International Conference on Computational Linguistics*, Stroudsburg, PA: Association for Computational Linguistics, pp. 2145-2158.

Zhou, G., Su, J., Zhang, J., and Zhang, M. 2005. "Exploring Various Knowledge in Relation Extraction," in *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*. Stroudsburg, PA: Association for Computational Linguistics, pp. 427-434.

Zhou, G., Qian, L., and Fan, J. 2010. "Tree Kernel-Based Semantic Relation Extraction with Rich Syntactic and Semantic Information," *Information Sciences* (180:8), pp. 1313-1325.

Zhou, P., Qi, Z., Zheng, S., Xu, J., Bao, H., and Xu, B. 2016. "Text Classification Improved by Integrating Bidirectional LSTM with Two-Dimensional Max Pooling," *arXiv:1611.06639*.

Zimbra, D., Abbasi, A., Zeng, D., and Chen, H. 2018. "The State-of-the-Art in Twitter Sentiment Analysis: A Review and Benchmark Evaluation," *ACM Transactions on Management Information Systems*, 9(2), article no. 5.