ESTABLISHING NOMOLOGICAL NETWORKS FOR BEHAVIORAL SCIENCE: A NATURAL LANGUAGE PROCESSING BASED APPROACH

Completed Research Paper

Jingjing Li
University of Colorado at Boulder
419 UCB, Boulder, CO80309
Jingjing.Li@colorado.edu

Kai R. Larsen
University of Colorado at Boulder
419 UCB, Boulder, CO80309
Kai.Larsen@coloradu.edu

Abstract

As the accumulated research base of the behavioral sciences have grown, the amount of actual knowledge discovery has not kept pace as evidenced by an increasing number of disconnected theories and the related problem of construct proliferation. Therefore, integrating social and behavioral sciences across research areas or even disciplines in a meaningful way is imperative. Despite the information systems (IS) discipline's leadership on creating nomological networks and inter-nomological networks for research integration, a quantitative approach to automatically establish nomological networks from large-scale data is missing. Based on the design science paradigm, we therefore propose a novel natural language processing based approach bringing together these two previous research endeavors. We used a dataset consisting of all the relevant behavioral studies from two tops journal in the IS and psychology fields to evaluate our approach in comparison to human decisions. Finally, the limitations and possible extensions of our approach are critically discussed.

Keywords: Natural language processing, Nomological Network, Behavioral science

Introduction

Human behaviors play a leading role in many critical areas including adoption of information systems, treatment of diseases, and achievement of education. During the past decades, the research community has seen strong growth of research and evidence on social behaviors, signified by a rapid accumulating of theories in social and behavioral sciences. For instance, Lee et al. (2004) shows over two hundred theories being used in information systems research, and Straker (2008) lists over three hundred theories or models that have some bearing on persuasion alone. Some of these theories have sparked thousands of extensions (e.g., the Theory of Planned Behavior – 11,655 citations; the Theory of Reasoned Action – 12,585 citations; and the Technology Acceptance Model – 9,200 citations: retrieved from Google Scholar on May 1, 2011).

Despite the net positive research growth in volume, the amount of actual knowledge discovered in social and behavioral science has not kept pace. Actually, the large number of theories has lead serious problems to some area of the behavioral sciences. For example, IS research in particular is now being characterized as theoretically scattered (Kraemer and Dutton 1991; Orlikowski and Baroudi 1991), fragmented (Banville and Landry 1989), and chaotic (Marble 2000). If researchers are unaware of closely related constructs or unable to validate their level of relatedness, behavioral theories will likely continue to grow apart, leading to overlapping or identical constructs that are seldom cited or reused, which we call as the *construct proliferation problem*.

For instance, Larsen's (2003) found that in one research area, 83 unique constructs were measured using 948 different scales. Most of the research papers employing these scales did not build on the existing similar scales, but rather relied on creating new ones. In addition, Colquitt & Zapata-Phelan (2007) identified two notable trends related to construct proliferation: many constructs have limited impact to management literature indicated by very few citations; the rate of new and reconceptualized constructs being introduced is increasing in recent years because of the Academy of Management Journal's recent explicated emphasis on interestingness, innovativeness and novelty in management research (Colquitt and Zapata-Phelan 2007). Behind these reasons may be the fact that behavioral science is pluralistic, different subgroups, even within a discipline, adopt different research paradigms, and new findings in one sub-discipline do not necessarily translate into changes in another (Cronbach 1987a, p.1297). Further, Furnas et al. (1987) found that different people are less than 20% likely to express the same idea using the same words, thereby making validation of the "sameness" of conceptualizations difficult, hampering scientific progress within the behavioral sciences. Consequently, due to this inconsistent language use across disciplines and pluralistic features of behavioral sciences, many researchers simply accept that, regardless of the topic of focus, it is impossible to find and incorporate all related published research.

In order to foster discovery of real innovative knowledge, it is imperative that we develop novel tools to integrate scattered constructs, fragmented theories and disconnected research within and across disciplines. In our context, we use the construct as the basic unit for analysis. As defined by Cronbach (1971, p. 464), a construct is "an intellectual device by means of which one construes events. It is a means of organizing experience into categories." Once found or developed by behavioral researchers, these constructs are then woven into theories – or nomological networks – "the interlocking system of laws which constitute a theory" (Cronbach and Meehl 1955, p. 290). A nomological network represents constructs as nodes, while every edge (connecting constructs) represents a relationship between constructs in a hypothesis. Cronbach (1987a; 1987b) made clear that measures of constructs are always open for interpretation, and must be interpreted in the context of their immediate nomological networks. Therefore, building a nomological network for construct interpretation could further our understanding of the constructs, thus alleviate the construct proliferation problem. In addition, Larsen et al. (2010) proposed an Inter-nomological-network (INN) infrastructure, which innovatively uses Latent Semantic Analysis (Deerwester et al. 1990) to discover synonymous constructs across paper or even discipline boundaries. However, while both nomological network and inter-nomological network offer opportunities to integrate constructs and theories in behavioral science, neither of them provides a quantitative approach to automatically extract nomological networks from behavioral studies. Although Larsen et al. (2010) took the first move towards automatically detecting construct similarities, the constructs and their co-occurrences (correlations) from hypotheses/propositions were manually collected. Therefore, we propose a novel natural language processing (NLP) based approach to address this research gap. In addition, we demonstrate the applicability by using a dataset consisting of relevant behavioral studies from MIS Quarterly—a top journal in IS field and Journal of Applied psychology—a top journal in psychology field, and evaluate the performance by comparing to human decisions. While two journals arguably represent a small sample size, the goal is to prove feasibility rather than to generalize.

We adopt the design science paradigm and in particular the guidelines of conducting design science by Hevner et al. (2004). Specifically, we follow the six activities of "a nominal process model for the conduct of design science research", which is introduced by Peffers et al. (2008) and is based on the guidelines by Hevner et al. (2004). Consequently, after the introduction section which discusses the relevance and motivation of the proposed problem (activity 1: "problem identification and motivation"), we specify the problem context by introducing relevant research on generating nomological network and inter-nomological network and identify the research gap (activity 2: "define the objectives for a solution"). Then, we develop our artifact as a novel natural language processing approach for identification and integration of constructs and theories from behavioral science (activity 3 "design and development"). In the forth section, we illustrate the applicability of the artifact by using a dataset comprised of articles from MIS Quarterly—a top journal in IS field and Journal of Applied Psychology—a top journal in psychology field (activity 4: "demonstration"). In addition, the performance of the artifact is evaluated (activity 5: "evaluation") by comparing with human decision. Finally, we summarize the results and discuss the future work in the final section (activity 6: "communication")

Related Work

Nomological Network

To explain the nomological network, we must understand two concepts: constructs and construct validity. Constructs are the cornerstone of the psychometric approach used throughout behavioral disciplines. As defined by Cronbach (1971, p. 464), a construct is "an intellectual device by means of which one *construes* events." In other words, a construct is a way for science to order observations. In behavioral research, a construct is generally some postulated attributes of people assumed to be reflected in test performance. Usually, researchers will create a measurement instrument (scale) consisting of a set of questionnaire items to operationalize a construct.

Construct validity (Cronbach and Meehl 1955, p. 290) refers to the degree to which a measurement actually reflects the theoretical construct. In other words, whether there is a match between the expected theoretical pattern and the observed/measured pattern. The construct proliferation problem is a type of construct validity issue in that researchers incorrectly treat measurements reflecting the same constructs as distinct. Since the 1950's a reasonably robust body of research has emerged establishing a set of standards to define and assess construct validity. The nomological network – "the interlocking system of laws which constitute a theory" (Cronbach and Meehl 1955, p.290)—was the first formal articulation of construct validity. A nomological network includes a theoretical framework representing the theoretical constructs and their relationships, an empirical framework demonstrating the measurements and their relationships, and the linkages between those two frameworks. In order to gain high construct validity, the intended measurement target must be clarified through examination of linkages residing in a nomological network. However, while the nomological network provides a philosophical foundation for construct validity, it does not provide a practical and applicable approach for actually assessing construct validity. An example of one such nomological network involving the affective commitment construct is presented in Figure 1, in which every arrow represents a hypothesis about a relationship between two constructs.

For the behavioral sciences, Cronbach (1987a; 1987b) made clear that measures of constructs are always open for interpretation, must be interpreted in the context of their immediate nomological network, and that their interpretation must be performed not by individuals but by the larger community of researchers. Unfortunately, no tool exists for the larger community to collaborate and validate constructs. The inter-Nomological Network (INN) infrastructure takes a first step toward addressing this limitation.

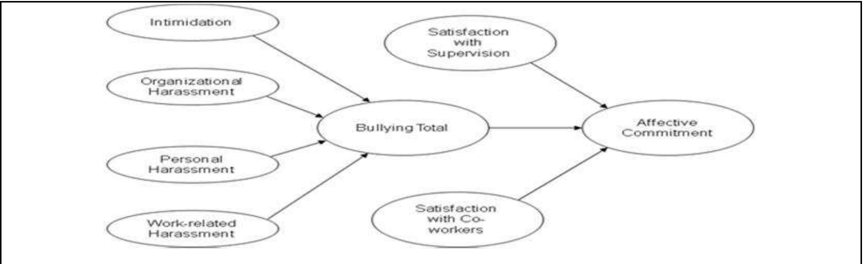


Figure 1. Example of a nomological network (McCormack et al. 2006).

Inter-Nomological Network

The inter-Nomological Network infrastructure (Larsen et al. 2010) is a combination of all the nomological networks in an area of a behavioral science. Several nomological networks are integrated by "synonymous" constructs presented in different theories and models. Latent Semantic Analysis (LSA) (Deerwester et al. 1990), which is a theory-based scientific method for extracting and representing the contextual-usage meaning of words by using Singular Value Decomposition (SVD), was used to detect these "synonymous" constructs.

The underlying idea of LSA is that the aggregate of all the word contexts, in which a given word does and does not appear, provides a set of constraints that determines the similarity of meaning of words, and sets of words, to each other (Landauer et al. 1998). Thus, when two terms occur in contexts of similar meaning – even in cases where they never occur in the same passage – LSA represents them as having similar meanings. (Landauer 2007). In fact, research has found that LSA can perform some meaning based tasks as well as humans, and LSA has been found to share up to 90% of the agreement between human experts (Landauer 2002). Latent Semantic Indexing (LSI) is a novel information retrieval approach, which uses LSA to capture the underlying semantics and improve search engine performance. It uses LSA to create a semantic space for a collection of document. An external text (a query) is then converted into a pseudo-document, and is projected as vectors into the semantic space. By calculating the semantic similarity (usually cosine similarity) between any pair of the query and the documents, LSI can retrieve a list of documents most semantically similar to the query. The primary tool used in INN is the extension of LSA and LSI, termed Stored LSI (sLSA), which works for analysis of text units down to the sentence-level and even word-level. The basic process is as follows. First, paragraph-level texts from academic articles containing construct definitions and questionnaire items were extracted to create an appropriate semantic space (an LSA high-dimensional space enabling context-specific language understanding); Next, constructs along with their definition and questionnaire items were collected by carefully trained research assistants. Subsequently, construct definitions and questionnaire items were converted into pseudo-documents (Deerwester et al. 1990), and were projected as vectors into the semantic space, enabling the similarities between any pair of questionnaire items or construct definitions to be represented as cosines. In the current version of INN, for two constructs being "synonymous", the cosines of at least two pairs of questionnaire items between these two constructs should be higher than a minimum cutoff cosine (which is defined by researchers). The formula below shows how to calculate the cosine similarity for two vectors v_1 and v_2.

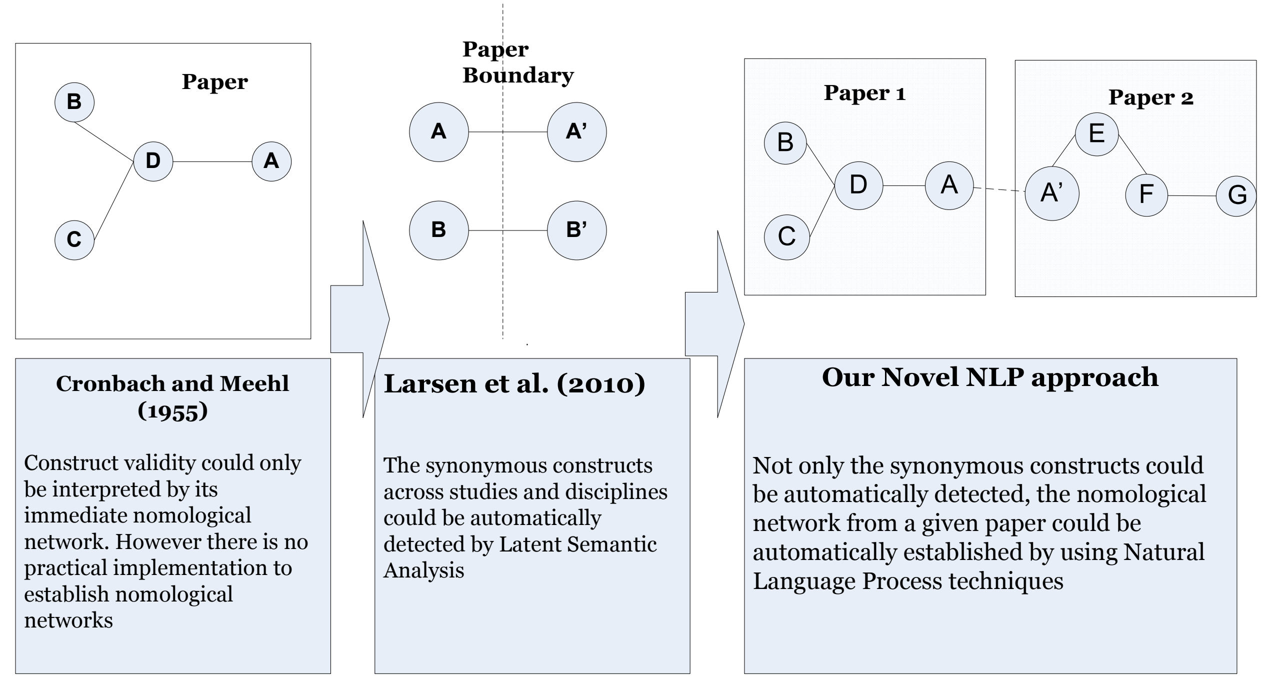
$$\text{Cosine similarity} = \frac{v_1 \cdot v_2}{\| v_1 \| * \| v_2 \|} \qquad (1)$$

Following this method, connections between synonymous constructs across nomological networks or disciplines could be automatically identified, which dramatically lightens the cognitive load placed on behavioral researchers, who are expected to find all relevant literature and constructs from large volume of research before publishing a new finding.

Research Gap

While both nomological network and inter-nomological network provide opportunity to integrate behavioral research, to our best knowledge, a quantitative approach to automatically generate nomological networks from individual studies is missing. While Cronbach and Meehl's idea of nomological network provides a philosophical foundation for better interpreting constructs, they do not provide an effective approach to establish nomological networks, especially from a large volume of research. Although the tool used in INN could automatically detect synonymous constructs across paper or even discipline boundaries, the process of extracting constructs and their relationships are done manually, which is labor-intensive and is also not appropriate for a large-scale data. Therefore, taking advantage of the latest NLP techniques, we propose a novel approach to automatically extract constructs and establish nomological networks from academic articles in behavioral sciences. Notice that in this paper we use the term "nomological network" to refer to a theory or a model in an individual study (Straub et al. 1995) and inter-nomological network to denote the ideas that integrate various theories or models across studies and disciplines.

Figure 2 demonstrates the contribution of our novel NLP-based approach. Our NLP approach seeks to automatically extract nomological networks from individual studies. Combining this approach with the inter-nomological network infrastructure proposed by Larsen et al. (2010), we could potentially integrate numerous theories and models in behavioral sciences thus providing a solution for construct validity and proliferation problem. Specifically, using LSA, a pair of synonymous constructs was detected (dotted lines), which serves as the "joint" point between these two nomological networks. Once these two nomological networks were integrated into one inter-nomological network, we could potentially obtain more relationships among constructs, e.g. construct A is likely to be the mediator between construct D and E. Researchers could test this hypothesized relationship later by conducting an empirical study. Once this relationship is validated, researchers could potentially combine two theories into one theory.



Cronbach and Meehl (1955)

Construct validity could only be interpreted by its immediate nomological network. However there is no practical implementation to establish nomological networks

Larsen et al. (2010)

The synonymous constructs across studies and disciplines could be automatically detected by Latent Semantic Analysis

Our Novel NLP approach

Not only the synonymous constructs could be automatically detected, the nomological network from a given paper could be automatically established by using Natural Language Process techniques

Figure 2 The ultimate goal of our novel NLP-based approach

Following this approach, a number of nomological network (theories) spreading across different disciplines could be integrated as a huge nomological network available for all researchers, thus promoting the cross fertilization of ideas and eliminating silos.

Novel Natural Language Processing Approach

Overall Framework

In order to automatically extract the nomological networks from publications in behavioral sciences, we propose a novel NLP-based approach composed of four steps. We will discuss the underlying assumption first.

We postulate that hypotheses, especially in quantitative and positivist type of research, usually represent the (or part of the) novel theory that authors want to validate, thus represent the (or part of the) nomological network. Because behavioral hypotheses usually include a statement describing the relationships between constructs, which equates a linkage connecting two nodes in the focal theory. Therefore, the goal of this study is automatic extraction of hypotheses, constructs, and their relationships from academic articles, as well as the provisioning of this evidence to researchers at large. In our system, we will also collect other types of variables, for instance demographic variables, which are not constructs. So we use "variable" to represent constructs and all other types of variables in the following sections.

We propose four steps to finish this process. Figure 3 explicates the process flow. First, academic articles on behavioral sciences were automatically downloaded from an online library portal. Second, hypotheses are extracted from the papers. Third, the variables existing in the hypotheses are extracted using information extraction methods. Finally, based on a paper's hypotheses along with the containing variables, a nomological network (or theory structure) for that paper could be automatically established.

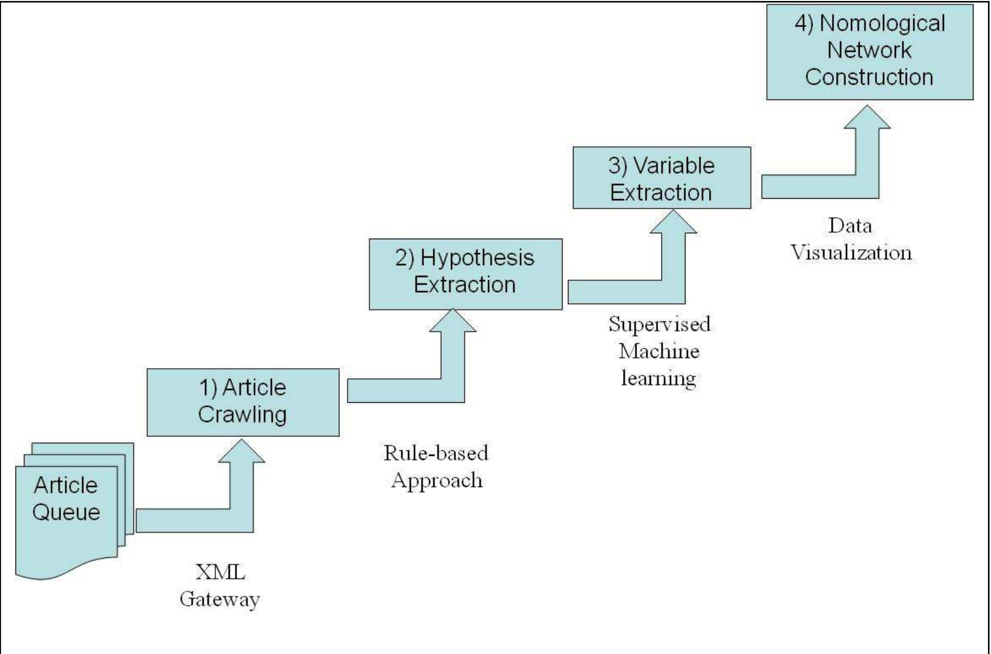


Figure 3 Overall framework for automatic establishment of nomological network

Step 1: Article Crawling

We follow a general web crawling approach to download html formatted articles from online academic libraries and parse them into plain text for later processing. The downloading process includes (1) URL construction (2) HTML tag cleaning (3) Meta information extraction. The way of constructing URL varies by different portals. In our demonstration, we used an XML gateway from ProQuest to construct the URLs which link to the web pages containing the full text of relevant articles.

Step 2: Hypothesis Extraction

The traditional rule-based information extraction approach is based on format and grammars. The format of hypotheses in social science research literature is generally quite prescribed and presents an opportunity to access the underlying knowledge in this literature. For example, in Venkatesh et al. (2000), several hypotheses addressed the influence of gender and social experience on the common constructs of the Technology Acceptance Model:

> 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.

Considering the task of differentiating those two hypotheses from sentences, a human annotator will first notice the following unique pattern:

> Pattern A: the hypothesis sentences are comprised of "a capitalized H + a number [+a letter] + a colon + a capitalized word + a string of words + a period".

This pattern then is converted to a regular expression (Kleene 1956). Regular expression is a language for specifying text search strings. It uses a formal language to represent the "patterns" or "rules" of a specific string. A search function will then search through the text based on this regular expression, and return all the matching texts.

```
‘H[0-9]+[a-zA-Z]?: [A-Z]\D+\.’
```

Subsequently, we could apply this rule to the remaining sentences in that article and only extract those satisfying sentences as hypotheses. Follow this manner, more formatting rules could be added to our rule system. However, as the complexity and the size of the rules system increases, the performance of this approach will likely decrease, which is known as a major issue of rule-based system. Therefore, a training data containing a subset of correctly identified hypotheses needs to be constructed as a gold standard. Through comparing the extraction results to the gold standard, we removed those rules that are overfitting to the hypothesis outliers.

Step 3: Variable Extraction

With a large database of hypotheses, variables in the form of single or multi-word phrases are available, but still need to be extracted. In the Venkatesh et al. (2000) example above, H1 contains three variables; perceived usefulness, behavioral intention, and gender (men vs. women). In general, to fulfill this variable extraction task, two NLP techniques are usually adapted, rule-based approach and supervised machine learning.

The rule-based approaches somewhat resembles the methods proposed for hypothesis extraction. With a carefully generated dictionary consisting of common variable names, a set of variable-matching rules could be derived. Those rules are then used to search through hypothesis text to identify unambiguous variable mentions. Whereas the rule-based method would achieve high precision with a sufficiently large dictionary, the precision drops dramatically when the dictionary is small. However, due to the practice of constantly renaming and reinventing variable names, there are structural limits on the success rate that may be expected from the rule-based approach.

Therefore, we adapt a supervised machine learning approach. This approach converts variable extraction problem into a classification problem: for each word in a sentence, we need to determine whether it belongs to a variable or not. Usually, we use a *label set* to represent this decision. A sample label set could be "Y" (is variable) and "N" (not a variable). In later section, we will introduce one of the commonly used label sets—IOB tags. Meanwhile, each word in the sentence has certain associated "features" which could differentiate it from other words and could facilitate the label assignment decision. We call these associated features as *input features*. Therefore, the classification problem is essentially how we can assign the correct label for a word given a set of its input features.

Typically, we use an inferred function (or classification function) to assign the class label. This inferred function needs to be generated through a training data set, because it helps to estimate some of the function parameters. A training set is comprised of a set of training examples, each of which is a pair consisting of several input features and a desired class label. Therefore, the variable extraction process can be divided into two steps: constructing training data and generating classification function, as shown in Figure 4:

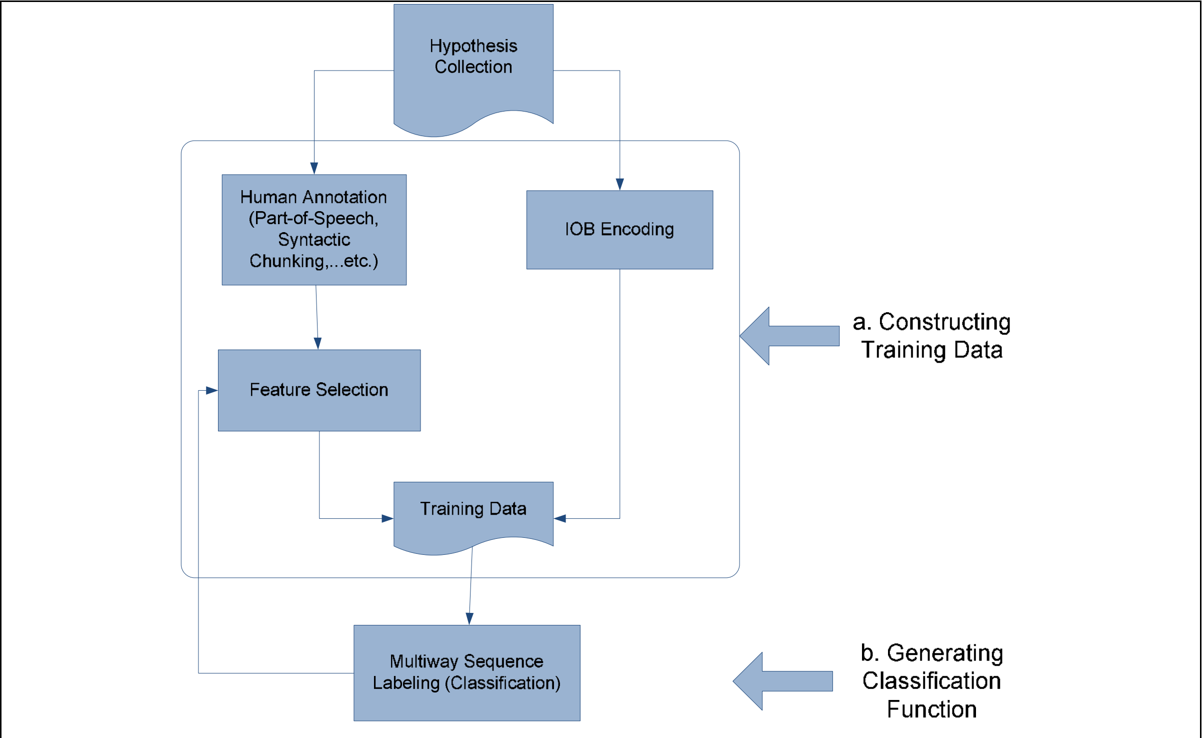

Hypothesis Collection
Human Annotation (Part-of-Speech, Syntactic Chunking,...etc.)
IOB Encoding
Feature Selection
Training Data
Multiway Sequence Labeling (Classification)
a. Constructing Training Data
b. Generating Classification Function

Figure4. Variable extraction process overview: Supervised machine learning

The first step is to construct a training data set. As mentioned before, each training example consists of two parts: input features and class label. The *IOB Encoding*, which is manually done by domain experts, allows each word to have a class label, indicating whether it belongs to a variable or not. On the other hand, the *Human Annotation* identifies plausible linguistic features of each word. The standard input features employed in state-of-the-art machine learning approaches include lexical, syntactical and shape features, such as parts-of speech, affixes, lemmas, orthographic pattern, etc. The number of input features for a certain word could be as many as hundreds or thousands in some sophisticated approaches. However, before generating the classification function we need either use the linguistic intuitions or apply some statistical measures to determine which features are most predictable. This is called *Feature Selection*. In addition, after we have created the classification function, we need to again evaluate which input features have improved the system performance the most so as to remove the noisy or useless features and further refine our feature set. After obtaining a training set, the second step is to generate a classification function. In our study, a state-of-art method—Conditional Random Fields (CRF) was applied to fulfill this task. In addition, the performance and generalizability of the classification function was evaluated by a test data which is separate from the training data but also contains the pairs of input features and IOB tags. The details are discussed in the following.

a. Constructing training data

We first introduce the IOB label set representing the occurrence of variable in hypotheses. Here, I is used to label tokens (words) inside of a variable, B is used to mark the beginning of a variable and O labels tokens outside any variables of interest. For instance, the hypothesis H1 in Venkatesh et al. (2000)

contains three variables, perceived usefulness, behavioral intention, and gender (men vs. women). After IOB encoding, the sentence will be:

> 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 influence/O women/B.

We then select a set of input features associated with each input word (i.e. each of the tokens to be labeled) to predict the IOB tags. These features should be plausible predictors of the IOB label, should be easily and reliably extractable from the source text and should also be related to domain knowledge. For example, the simplest input feature for a word is itself (also called as token), because a word has its unique spelling. Once an adequate set of features is chosen, we could encode them into an appropriate form for future processing. For example, in Venketesh et al. (2000), variables are often noun phrases. In linguistics, we call a group of words behaving like a single unit or phrase as "syntactic chunk" or "constituency". After using syntactic chunking (i.e. domain experts manually identify the syntactic chunks for each sentence) to identify all the noun phrases of the sentence, the first hypothesis in Venkatesh et al. (2000) will take the following form:

> [$_{NP}$ Perceived usefulness] will influence [$_{NP}$ behavioral intention] to use a [$_{NP}$ system] more strongly for [$_{NP}$ men] than [$_{NP}$ it] will influence [$_{NP}$ women].

Routinely, the presence of a syntactic chunk in training set is also represented in IOB forms. In order to differentiate it from our variable class label, we use B_{NP} to indicate the beginning of a noun phrase, B_{VP} to indicate the beginning of a verb phrase, and so on so forth.

Table1 illustrates a training example using *tokens*, *part of speech* and *syntactic chunk* (TOK+POS+CHUNK) as input features.

Table 1 Training examples for supervised machine learning

Input Features			**Class Label**
Token	*Part of Speech*	*Syntactic chunk*	*IOB tag*
Perceived	VBN	B_{NP}	B
usefulness	NN	I_{NP}	I
will	MD	B_{VP}	O
influence	VB	B_{VP}	O
behavioral	JJ	B_{NP}	B
intention	NN	I_{NP}	I
to	TO	B_{VP}	O
use	VB	I_{VP}	O
a	DT	B_{NP}	O
system	NN	I_{NP}	O

b. Generating classification function

The last step is to choose an appropriate algorithm to train a classification function (classifier) based on training data. Algorithms such as Conditional Random Fields (CRF), Maximum Entropy Markov Models (MEMMs), Hidden Markov Models (HMMs), and Support vector machines (SVMs) all fulfill the requirements for this task. In this paper, we applied Conditional Random Fields (Lafferty et al. 2001), which is a probabilistic model for labeling sequential data and has been shown to outperform generative models such as the Hidden Markov Model (HMMs) and Maximum-Entropy Markov Model (MEMM). Figure 5 shows the graphical representation of a linear-chain CRF.

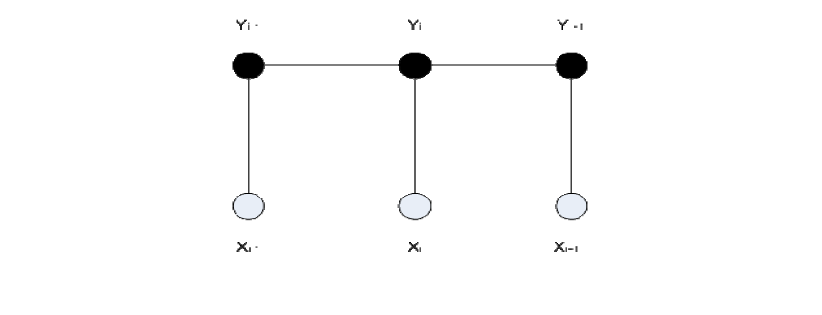


Figure5. Graphical representation of a linear-chain CRF

According to Figure5, a sequence labeling task is then converted to assign a *optimal* label sequence (e.g. the IOB tags) $Y = (y_1, y_2, \ldots y_n)$ to a set of observation sequence (e.g. the input features, such as tokens, POS tag, etc.) $X = (x_1, x_2, \ldots x_n)$. In this case, CRF uses a conditional probability to assess the extent of this "optimization", i.e. the IOB sequence with the highest conditional probability will be the optimal labeling solution. The formula to calculate this conditional probability is:

$$P(\overrightarrow{y} \mid \overrightarrow{x}) = \frac{1}{Z_{\vec{x}}} \exp(\sum_{i=1}^{n} \sum_{k} \lambda_k f_k (y_{i-1}, y_i, \vec{x}, i)) \quad (2)$$

In this formula, k stands for the number of feature functions. The feature functions $f_k (y_{i-1}, y_i, \vec{x}, i)$ transform input features into a set of real values (Wallach 2004). Z is a normalization factor. λ is a weight corresponding features and needs to be estimated from training data. For each input token x_i, three conditional probabilities $P(I \mid x_i)$, $P(O \mid x_i)$ and $P(B \mid x_i)$ corresponding to I, O and B labels are computed respectively. A dynamic programming algorithm, usually Viterbi, A* search or the combination of these two (Lafferty et al. 2001) , is then employed to search for the optimal IOB sequence based on all the previously computed probabilities. More details of this algorithm could be found in Lafferty et al. (2001).

Step 4: Nomological Network Construction

Having detected the variables within each hypothesis, we detect reuse of variables across hypotheses within a given paper. By treating each unique variable as a node and the relationship between variables as a linkage (this linkage only represents the undirected correlation), a network for that specific paper is generated as in Figure 6[1].

Since researchers tend to use similar phrases to represent variables within a paper, this task is relatively easy to fulfill. For those variables of slightly different names within different hypotheses, we adopted edit distance (also called Levenshtein distance) to assess the phrase similarity (Levenshtein 1966). The edit distance between two string sequences is defined as the minimum numbers of edit operations needed to transform one string to another. Each of the three operations—insertion, deletion and substitution has a cost of 1. A dynamic programming called minimum edit distance algorithm (Wagner and Fischer 1974) is then used to compute the minimum number of edits between two string sequences. Finally, we use a minimum edit distance ratio ranging from zero to one to indicate the string similarity, which is the minimum edit distance divided by the maximum length of two variables. We use a cutoff point of this ratio to determine whether an integration of two variables within a paper should be made.

[1] Note that we only consider the co-occurring patterns of variables inside a hypothesis as denoted in linkages without directions. Therefore, this approach does not capture the moderation and mediation relationships in hypotheses containing more than two variables.

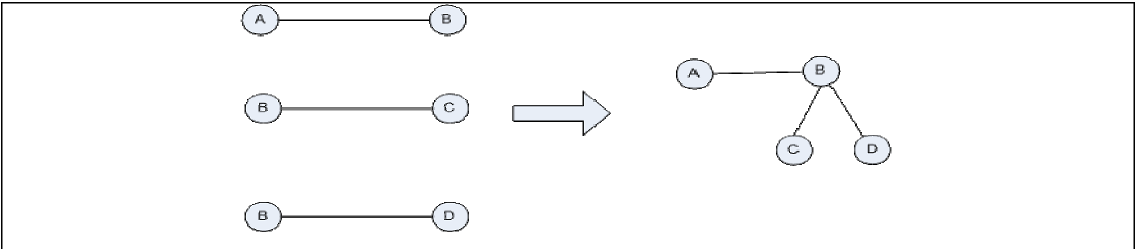


Figure 6. Building nomological network from constructs and hypotheses

Demonstration and Evaluation

To demonstrate and evaluate our NLP approach for automatic creation of nomological networks, we used 342 articles from MIS Quarterly (MISQ) and Journal of Applied Psychology (JAP). While the sample size is relatively small, our main goal is to demonstrate feasibility rather than generalize. In addition, since we aim to compare our automatic approach to manual process, a small sample size is thus adequate for human experts to handle. Certainly, one of the major advantages of our NLP approach is to address large scale dataset within and across disciplines of behavioral sciences.

In this section, we first illustrate the applicability of our NLP approach. Then we evaluate its performance in comparison to human decisions. Finally, we highlight and discuss limitations of our NLP approach.

Application of NLP approach

Step 1: Article crawling

We adopted the web crawling techniques to download all the articles in MISQ between 1991 and 2005 and JAP between 1998 and 2007 from ProQuest. All papers were carefully examined by a senior faculty member, and any paper containing at least on construct and one hypothesis was selected for inclusion into the study, for total of 122 articles in MISQ and 220 articles in JAP. For each article, the reference information including author, title, publication year, journal, ISSN was automatically extracted and stored in database. The full text of each included article was converted to plain text by using regular expressions (Kleene 1956) to remove HTML tags.

Step2: Hypothesis Extraction

In order to create a set of formatting rules, we first created a training data set. We randomly selected 60 articles with 42,575 sentences and had research assistants manually identify the 443 hypotheses contained within the articles. This dataset containing 443 hypotheses is used as a gold standard. We then started with a preliminary rule, and used regular expression to extract hypotheses from the same 60 articles. Through comparing the extraction result to the gold standard, a set of formatting rules were iteratively added or removed. We used precision, recall and F measure, the traditional information extraction metrics, to demonstrate the performance of this rule-based approach. Precision deals with the fraction of correctly identified hypotheses among those hypotheses that the algorithm extracted, which also equates one minus false positive rate (Type I error rate). Recall deals with the fraction of correctly identified hypotheses among all the hypotheses existing in the articles, which also equates one minus false negative rate (type II error rate).

Just like type I error and type II error, there is a tradeoff between precision and recall: One can increase precision by extracting fewer hypotheses thus decreasing the recall, or boost recall by extracting more hypotheses but at the expense of low precision. Practically, precision is usually emphasized more. We stopped modifying the formatting rules when the precision reached 91.11% along with a reasonably high recall (89.40%).

In order to obtain a balanced view of precision and recall, we further used F measure, which is a harmonic mean of precision and recall (van Rijsbergen 1975). Therefore, the F measure for the training data is 90.25%.

$$F\ Measure = \frac{2PR}{P+R} \quad (3)$$

We then tested our rule-based approach on a test data, which is separate from the training data set described above. In order to eliminate the random errors, we sampled 60 articles from all the articles 4 times, and had research assistants manually identified each of the gold standards. We then applied our rule-based approach on the 4 test datasets, and obtained the average precision, recall and F measure, which indicates the true system performance. The results were shown in Table 2.

Table 2 Performance of rule-based hypothesis extraction approach

Measurement	**value**	**Parsing Time**[2]
Precision	85.39%	104.31 Seconds
Recall	84.64%	
F Measure:	84.99%	

Finally, we applied this approach to all 342 articles, and extracted 2016 hypotheses from 209,634 sentences with the total running time less than 2 minutes. The extracted hypotheses are further examined by research assistants and resulted in a total of 2050 hypotheses for variable extraction.

Step3: Variable Extraction

In order to evaluate the performance of the variable extraction system, we hired several research assistants to annotate the existence of constructs among those 2,050 hypotheses extracted in step2 and found 5,769 variable incidences and 2,581 unique variables. Among which, 930 hypotheses are from MISQ, with an average of 2.49 variables per hypothesis, and the rest 1,120 hypotheses come from JAP, with an average of 3.08 variables per hypothesis.

This dataset was then divided into training and test data sets. In order to obtain a robust estimate of the algorithm, we applied a 5-fold cross validation. Specifically, we first randomly partitioned the dataset into five equally large (410 instances) and complementary subsamples. We then picked one subsample as the test data at a time, and combined the rest 4 subsamples as a training data set, which produced five different pairs of training and test sets. Therefore, the model performance is calculated as its average performance across the five test sets.

As we described earlier, the input features for training data should be plausible predictors of IOB tags and need to be carefully selected based on domain knowledge. Therefore we employed four different sets of input feature to train the CRF model. The four feature sets are tokens (TOK), tokens + part of speech tags (TOK+POS), tokens +syntactic chunks (TOK+CHUNK), and tokens + part of speech tags + syntactic chunks (TOK + POS + CHUNK). We adopted 45 part of speech tags from Penn Treebank (Marcus et al. 1993) and 3 syntactic chunks—noun phrases (NP), verb phrases (VP) and proposition phrases (PP) from the shared task in CoNLL-2000. The selection of best input feature set is based on the related model performance.

We compared the extracted variables with human annotated gold standard, and represented the evaluation results as *precision*, *recall* and *F measure*. In addition, we also used *accuracy* to specifically

[2] CPU Intel (R) Core2Duo 2.67GHz, Ubuntu, 8G RAM.

evaluate the IOB tagging performance, which equals the number of corrected identified IOB tags divided by the number of all IOB tags in the data set.

In order to demonstrate the effectiveness of the system we compared the accuracy rate with a random guess baseline, which is the percentage of highest occurring tags in the test data (the "O" tag in this case).

The model performances in terms of four feature sets and four evaluation metrics are shown in Table 3. The average parsing time for CRF to extract variables from five different tests set is 151.8 seconds which is far less than human annotation.

Table 3 Variable extraction performances (Boldface: best performance for a given setting)

	TOK	TOK+POS	TOK+CHUNK	TOK+POS+CHUNK
Precision	74.10%	**74.65%**	73.09%	73.66%
Recall	68.14%	**69.59%**	68.07%	68.65%
F Measure	70.98%	**72.02%**	70.47%	71.05%
Average Accuracy	89.50%	**89.92%**	89.20%	89.74%
Accuracy Baseline (% of tag O)	**74.14%**			

In general, our approach with all the feature sets outperforms the accuracy baseline by around 15%. The feature set with tokens and part of speech tags (TOK+POS) surpassed the rest configurations. For all the variables extracted by using this feature configuration, 74.65% agreed with expert decisions. Meanwhile, this configuration covered 69.59% of the variables identified by domain experts. The IOB tagging accuracy is 89.92% which is 15.78% higher than the random guess model. The feature set with tokens, part of speech tags and syntactic chunks (TOK+POS+CHUNK) performs the second best (F measure: 71.05%). Notice that adding syntactic chunk information to the feature set actually hurt the system performance, as shown in the poor performance with TOK+CHUNK configuration. Hence, the good performance of feature set TOK+POS+CHUNK is actually due to the inclusion of part of speech tags (POS) but not the combination of POS and CHUNK. Therefore, when sufficient feature information is not available, the best strategy is to only include tokens (TOK) as features, which, in this data set, gave the second best precision and a decent recall.

However, we still have plenty of room for performance improvement in the sense that we didn't employ other potentially useful feature such as shading feature (i.e. plural, single, capitalized, and lower) and we didn't combine this result with rule-based approach.

Step 4: Nomological Network Construction

After the extraction of hypotheses and variables from articles, we represented hypotheses with nodes and linkages, which stand for variables and their relationships respectively.

Assuming that variables contain in the same hypothesis are often related, we first connected variables within a single hypothesis. For a hypothesis with a direct relationship between two variables, we directly connect these two variables with an undirected link. However, since we only extracted variables from hypotheses without knowing the role of them, for a hypothesis with more than two variables--usually represented as moderation and mediation, we simply connect the variables in all possible ways. Certainly, this method will create many unwanted links, which is a more severe problem for JAP than MISQ. According to the previous section, MISQ has on average 2.49 variables per hypothesis, while JAP has 3.08. This indicates that JAP articles tested more mediation or moderation effect than MISQ. In order to give an estimate of how our variable-connecting approach performs, we randomly sampled 40 articles from each journal, and had domain experts identify the relationship type for every hypothesis. For all 232 hypotheses from JAP, 38.36% of them are testing moderation, mediation or direct relationship between more than two variables, while only 18.57% of 332 MISQ hypotheses are related to these relationship types.

In addition, we didn't differentiate empirically supported hypotheses from unsupported ones, since this statistical information usually is not represented in the hypothesis text. In order to estimate how disregarding empirical support information will affect our nomological network construction performance, we again randomly sampled 40 articles from each journal, and had domain experts carefully read through each article and classify the containing hypotheses into two classes: "supported" and "unsupported". The result shows that ignoring empirical support information is a more severe problem for MISQ than JAP, with 31% unsupported hypotheses in MISQ and only 23.7% in JAP.

Second, we combined hypotheses within one article through their shared variables. We used the minimum edit distance ratio to assess the similarity between variables from different hypotheses. We set the cutoff point as 0.95 so that for any pair of variables with minimum edit distance ratio greater than 0.95, we combined them as one variable. This cutoff point is relatively high because the textual representations for two different variables in one article tend to be highly similar. For instance, the minimum edit distance ratio between two different variables "Communication decoding competence" and "Communication encoding competence" in Ko et al. (2005) is 93.9%.

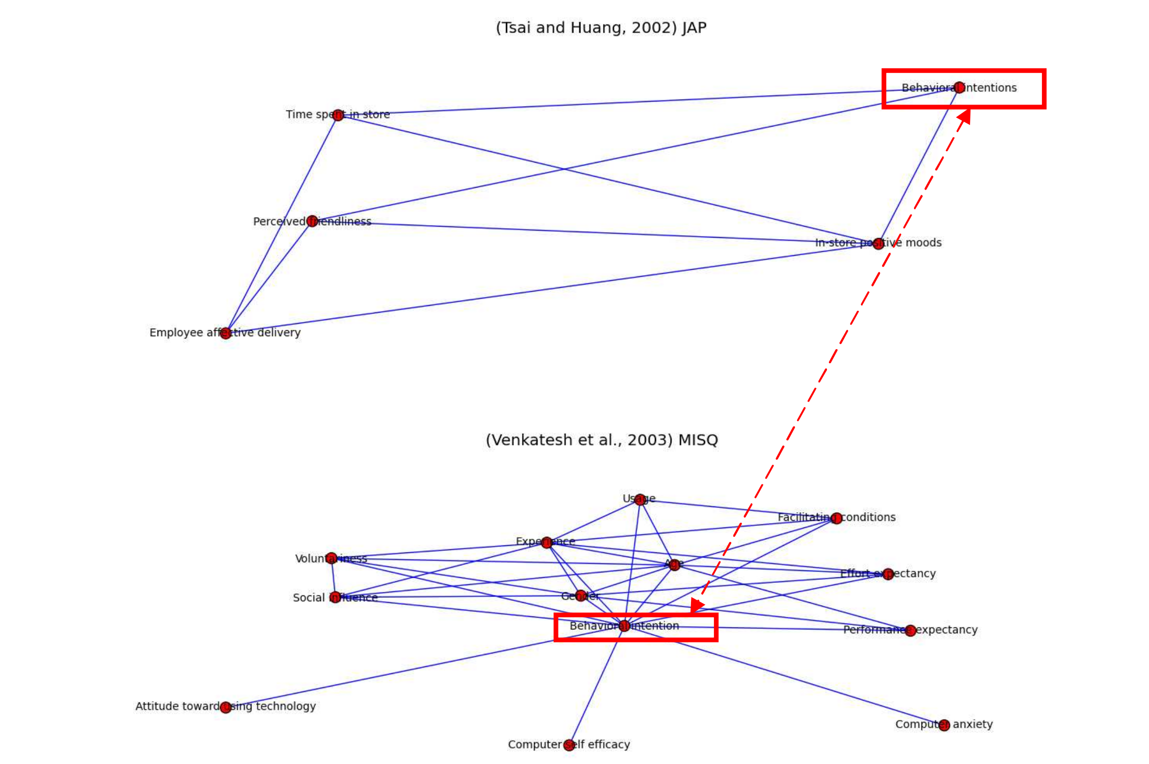


Figure 7. Four nomological network extracted using our approach

After connecting variables within and across hypotheses for each article, a nomological network is automatically established for this article. Figure 7 demonstrates two nomological networks extracted from Venkatesh et al. (2003) and Tsai & Huang (2002). Notice that the moderation relationships in Venkatesh et al. (2003)'s model are not represented correctly because of the simplistic variable-connecting approach we adopted. While Venkatesh et al. (2003) proposed a unified theory of information technology acceptance research, Tsai and Huang (2002) explored the relationship between employee affective delivery and customer reactions (e.g. customer satisfaction, word of mouth and purchase behavior). By looking at these two nomological networks, we argue that it is possible to connect the shared variables *behavioral intentions* in the future, even though those two variables may be operationalized differently.

By combining these two variables together, a number of research questions could be asked. For example, a potentially interesting research question is whether employee affective delivery is positively correlated with social influence.

Discussion and Limitations

While the theoretical implication of this study has already been highlighted, our NLP approach also has significant practical contribution. Assume that you are a behavioral researcher who wants to investigate the importance of *flexibility* in interpersonal relationships. A search for *interpersonal* and *flexibility* will list all papers containing those words in any database the researchers use. For example, Google Scholar (available at http://scholar.google.com) will list 142,000 papers, ranging from papers on cognitive flexibility theory to flexibility of sound. Our approach thus helps to avoid this deluge of information by narrowing the search down to the articles with *flexibility* as variables.

However, the initial version of this research depends on the existence of hypotheses in published research. Our research therefore omits variables existing in non-hypothesis research. The approach is limited to hypothesis research because hypotheses follow a highly prescribed format, which offers a superior starting point for NLP-based approach. However, the starting point focuses on quantitative, positivist research, potentially ruling out other types of behavioral variable research. Furthermore, our approach incorporates all the hypotheses into the nomological networks, disregarding the level of empirical support.

While extracting variables directly from hypotheses allows us to identify the existence of relationships, our approach does not extract the *direction* for these relationships, or in other words the causal relationships. Such causal relationships are critical when integrating nomological networks in the behavioral sciences, since knowing the hypothesized directionality would immediately allow a researcher to find a structural weakness in existing research. The alleviation of cognitive load that accompanies this information could also foster novelty and progress in behavioral science. Therefore, we are currently working on a hypothesis generation system which also extracts the causal relationships by combining a sentence constituency parser with a dependency parser.

Conclusion

Despite the explosive growth of the behavioral sciences, the actual knowledge discovered has not kept pace. A large number of theories have made some area of the behavioral sciences theoretically scattered, fragmented and chaotic forcing researchers to reinvent hypothesis relationships that are already introduced by others. While replication serves a key role in science, unintended replication leads to construct proliferation.. Therefore, integration of the behavioral sciences is highly demanded.

The “integration” problem existing in the behavioral sciences is analogous to the problem encountered by Chemistry 140 years ago:

> When scientists first tried to describe the physical and chemical properties of the elements and chemical compounds, which are formed by the combination of atoms of different elements, they soon became buried under a mountain of seemingly unconnected facts. Many early scientists recognized the need to organize this information, and they attempted to discover some sort of order or pattern that could simplify what seemed to them an overwhelming array of chemical facts. The solution to the problem was [Mendeleev’s] periodic table of the chemical elements” (Stwertka 2002, p.7).

It is clear that the behavioral sciences now exist in a “pre-Mendeleev” era, where human cognitive limitations require research areas to become narrower if they are to extend their depth. Therefore, there is a demand for identifying and extracting the “atomic” elements of behavioral science research, and then weaving them into some sort of reasonable pattern. By “atomic” elements we refer to these variables that are the cornerstone of the psychometric approach used throughout behavioral disciplines. By “pattern” we refer to these nomological networks that weave constructs into theories. Even though some previous studies (e.g., Larsen et al. 2010) have made some efforts to automatically integrate constructs and nomological networks, we argue that a quantitative approach for automatically extracting nomological network, which is especially useful with large-scale data, is missing.

We follow the design science paradigm introduced by Hevner et al (2004) to structure this paper. After the introduction of the problem and the identification of the research gap, we elaborated our novel NLP-based approach in four steps: paper downloading, hypothesis extraction, variable identification and nomological network extraction. To demonstrate the applicability of our approach, we used a dataset consisting of all the relevant behavioral studies from MIS Quarterly and Journal of Applied Psychology. We also compared the accuracy and the required time of our automatic approach with the human experts' manual annotation. Our preliminary results showed reasonable results and a huge potential left to unleash. We also critically discussed the limitations of our approach regarding underlying assumption and approach process. Specifically, our approach only considers those papers that contain hypotheses, and establishes nomological network based on both supported and unsupported hypotheses. Moreover, our approach does not take casual relationship between variables into account. However, we've been working on extracting the direction for variable relationships by using constituency parser and dependency parser. In future research, we intend to apply our approach to more journals in Information Systems, Management, Marketing and Psychology fields so as to provide a shared knowledge backbone for researchers in behavioral science.

Acknowledgement

We thank the U.S. National Science Foundation for research support under grant NSF 0965338 and the Colorado Clinical & Translational Sciences Institute for research support under NIH/CTSI 5 UL1 RR025780. We are grateful to Julia I. Lane and Randy Ross.

References

Banville, C., and Landry, M. 1989. "Can the Field of Mis Be Disciplined?," *Communications of the ACM* (32:1), pp. 48-60.

Colquitt, J.A., and Zapata-Phelan, C. 2007. "Trends in Theory Building and Theory Testing: A Five-Decade Study of the Academy of Management Journal," *Academy of Management Journal* (50:6), 12, pp. 1281-1303.

Cronbach, L.J. 1971. "Test Validation," in *Educational Measurement,* R.L. Thorndike (ed.). Washington, D.C.: American Council on Education, pp. 443-507.

Cronbach, L.J. 1987a. "Construct Validity after Thirty Years," in *Linn, R. L.,* T. Intelligence: Measurment, and Public Policy (ed.). Urbana, Ill.: University of Illinois Press.

Cronbach, L.J. 1987b. "Five Perspectives on Validity Argument," in *Test Validity for the 1990s and Beyond,* H. Wainer and H. Braun (eds.). Hillsdale, N.J.: Erlbaum.

Cronbach, L.J., and Meehl, P.E. 1955. "Construct Validity in Psychological Tests," *Psychological Bulletin* (52:4), pp. 281-302.

Deerwester, S., Dumais, S., Furnas, G., Landauer, T., and Harshman, R. 1990. "Indexing by Latent Semantic Analysis," *Journal of the American Society for Information Science* (41:391-407).

Furnas, G.W., Landauer, T.K., Gomez, L.M., and Dumais, S.T. 1987. "The Vocabulary Problem in Human-System Communication," *Communications of the ACM* (30:11), pp. 964-971.

Hevner, A.R., March, S.T., Park, J., and Ram, S. 2004. "Design Science in Information Systems Research," *MIS Quarterly* (28:1).

Kleene, S.C. 1956. "Representation of Events in Nerve Nets and Finite Automata," in *Automata Studies,* C.E. Shannon and J. McCarthy (eds.). Princeton University Press, pp. pp. 3-42.

Ko, D.-G., Kirsch, L. J. and King, W. R. (2005) *MIS Quarterly* **(**29**)**, 59-85.

Kraemer, K.L., and Dutton, W.H. 1991. "Survey Research in the Study of Management Information Systems," in *The Information Systems Research Challenge: Survey Research Methods,* K.L. Kraemer (ed.). Boston, MA: Harvard Business School, pp. 3-57.

Lafferty, J., McCallum, A., and Pereira, F. 2001. "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data," *18th International Conf. on Machine Learning*, Morgan Kaufmann, pp. 282-289.

Landauer, T.K. 2002. "On the Computational Basis of Cognition: Arguments from Lsa," in *The Psychology of Learning and Motivation,* B.H. Ross (ed.). New York, NY: Academic Press.

Landauer, T.K. 2007. "Lsa as a Theory of Meaning," in *Handbook of Latent Semantic Analysis,* D.S.M.

Thomas K Landauer, Simon Dennis, and Walter Kintsch (ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Landauer, T.K., Foltz, P.W., and Laham, D. 1998. "Introduction to Latent Semantic Analysis," *Discourse Processes* (25), pp. 259-284.

Larsen, K.R., Lee, J., Li, J., and Bong, C.H. 2010. "A Transdisciplinary Approach to Construct Search and Integration," in: *AMCIS 2010 Proceedings*. Lima, Peru: p. 524.

Larsen, K.R.T. 2003. "A Taxonomy of Antecedents of Information Systems Success: Variable Analysis Studies," *Journal of Management Information Systems* (20:2), Fall2003, pp. 169-246.

Lee, Y., Lee, Z., and Gosain, S. 2004. "The Evolving Intellectual Diversity of the Is Discipline: Evidence from Referent Theoretical Frameworks," *Communications of the AIS* (13), pp. 546-579.

Levenshtein, V.I. 1966. "Binary Codes Capable of Correcting Deletions, Insertions, and Reversals." Soviet Physics Doklady.

Marble, R.P. 2000. "Operationalising the Implementation Puzzle: An Argument for Eclecticism in Research and in Practice," *European Journal of Information Systems* (9:3), pp. 132-147.

Marcus, M.P., Santorini, B., and AMarcinkiewsicz, M.A. 1993. "Building a Large Annotated Corpus of English: The Penn Treebank.," *Computational Linguistics* (19:2), pp. 313-330.

McCormack, D., Casimir, G., Djurkovic, N., and Yang, L. 2006. "The Concurrent Effects of Workplace Bullying, Satisfaction with Supervisor, and Satisfaction with Co-Workers on Affective Commitment among Schoolteachers in China," *International Journal of Conflict Management* (17:4), pp. 316-331.

Orlikowski, W.J., and Baroudi, J.J. 1991. "Studying Information Technology in Organizations: Research Approaches and Assumptions," *Information Systems Research* (2:1), 03, pp. 1-28.

Peffers, K., Tuunanen, T., Rothernberger, M., and Chatterjee, S. 2008. "A Design Science Research Methodology for Information Systems Research," *Journal of Management Information Systems* (24:3), pp. 45-77.

Straker, D. 2008. *Changing Minds: In Detail*. Syque Press.

Straub, D., Limayem, M., and Karahanna-Evaristo, E. 1995. "Measuring System Usage: Implications for Is Theory Testing," *Management Science* (41:8), pp. 1328-1342.

Stwertka, A. 2002. *A Guide to the Elements*, (2nd edition ed.). New York, NY: Oxford University Press.

Tsai, W.-C., and Huang, Y.-M. 2002. "Mechanisms Linking Employee Affective Delivery and Customer Behavioral Intentions," *Journal of Applied Psychology* (87:5), pp. 1001-1008. van Rijsbergen, C. 1975. "Information Retrieval," *Butter-worths*.

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 Q.* (24:1), pp. 115-139.

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

Wagner, R.A., and Fischer, M.J. 1974. "The String-to-String Correction Problem," *Journal of the Association for Computing Machinery* (21), pp. 168-173.

Wallach, H. 2004. "Conditional Random Fields: An Introduction," *Technical Report*.