1-1-2014

Estimating the magnitude of method bias on account of text similarity using a natural language processing-based technique

Rajeev Sharma
University of Wollongong, rajeev@uow.edu.au

Murad Safadi
University of Wollongong, ms34@uowmail.edu.au

Megan Andrews
University of Wollongong, mea816@uowmail.edu.au

Philip O. Ogunbona
University of Wollongong, philipo@uow.edu.au

Jeff Crawford
University of Tulsa

Estimating the magnitude of method bias on account of text similarity using a natural language processing-based technique

Abstract

A number of potential biases have been identified that may contribute a spurious component to the observed correlation between variables. One such potential bias is the manner in which items are worded. This paper presents a technique and a program of research to estimate the magnitude of method bias arising from item wording. We hypothesize that the greater the level of text similarity between items employed to capture predictor and criterion variables, the greater the magnitude of the observed effect size between them. Two samples will be employed; one investigating the perceived usefulness-use correlations reported in the TAM literature and the other investigating the attitude behavior (physical activity) correlation reported literature drawing upon Theory of Planned Behavior (TPB). An NLP-based technique is developed to rate predictor-criterion pairs on similarity. The hypothesis will be tested by meta-regressing the predictor-criterion correlations against their respective similarity scores. Implications for research and practice are discussed.

Keywords

method, technique, magnitude, processing, estimating, language, natural, similarity, text, account, bias

Disciplines

Engineering | Science and Technology Studies

Publication Details

Sharma, R., Safadi, M., Andrews, M., Ogunbona, P. O. & Crawford, J. (2014). Estimating the magnitude of method bias on account of text similarity using a natural language processing-based technique. 35th International Conference on Information Systems "Building a Better World Through Information Systems", ICIS 2014 (pp. 1-10). AIS Electronic Library.

Estimating the magnitude of method bias on account of text similarity using a natural language processing-based technique

Submission Type: Research-in-Progress

Rajeev Sharma
University of Wollongong
Wollongong NSW 2522
rajeev@uow.edu.au

Murad Safadi
University of Wollongong
Wollongong NSW 2522
murad@uow.edu.au

Megan Andrews
University of Wollongong
Wollongong NSW 2522
megan_andrews@uow.edu.au

Philip Ogunbona
University of Wollongong
Wollongong NSW 2522
philipo@uow.edu.au

Jeff Crawford
Lipscomb University
1 University Park Drive
Nashville TN 37204-3951
jeff.crawford@lipscomb.edu

Abstract

A number of potential biases have been identified that may contribute a spurious component to the observed correlation between variables. One such potential bias is the manner in which items are worded. This paper presents a technique and a program of research to estimate the magnitude of method bias arising from item wording. We hypothesize that the greater the level of text similarity between items employed to capture predictor and criterion variables, the greater the magnitude of the observed effect size between them. Two samples will be employed; one investigating the perceived usefulness-use correlations reported in the TAM literature and the other investigating the attitude-behavior (physical activity) correlation reported literature drawing upon Theory of Planned Behavior (TPB). An NLP-based technique is developed to rate predictor-criterion pairs on similarity. The hypothesis will be tested by meta-regressing the predictor-criterion correlations against their respective similarity scores. Implications for research and practice are discussed.

Keywords: Method bias, item similarity, method-method pair technique, NLP, Jaccard similarity measure, semantic similarity

Introduction

The potential of methodological artifacts to bias the findings of empirical studies has long been acknowledged (Cook and Campbell 1979; Podsakoff et al. 2003; Podsakoff et al. 2012). An extensive literature has described the potential sources of method bias, the mechanisms through which those biases operate, and recommendations for designing empirical studies to prevent or minimize the extent of method effects (Burton-Jones 2009; Podsakoff et al. 2003; Podsakoff et al. 2012).

The literature has also proposed a number of techniques to estimate and control for the magnitude of method bias. These include the Multitrait Multimethod (MTMM) technique, the Harman one-factor test, the marker variable technique, and the Unmeasured Latent Method Construct (ULMC) technique, among others (Avolio et al. 1991; Campbell and Fiske 1959; Chin et al. 2012; Lindell and Whitney 2001). The MTMM technique is often referred to as the 'gold-standard' technique as, unlike the other techniques, it relies on variability in methods to estimate and partial out the effect of method bias. However, it is rarely employed in practice. The other techniques are characterized by two key features – within-study estimation of the magnitude of method bias and estimating the effect of method without any variability in method. Empirical assessments of the other techniques have raised strong concerns about the validity and reliability of those techniques (Chin et al. 2012; Richardson et al. 2009; Sharma et al. 2004; Sharma et al. 2010; Sharma et al. 2009; Straub and Burton-Jones 2007).

Departing from the tradition of within-study estimation of the magnitude of method bias, Sharma et al. (2009) proposed the method-method pair technique to estimate the magnitude of method bias based on between-study variation in methods. Based on theoretically derived expectations, Sharma et al. (2009) developed a schema to rank the relative magnitudes of bias in correlations that could be expected when specific method-method pairs are employed to measure the constructs involved in a correlation. Sharma et al. (2009) illustrated the technique by analyzing the cumulative empirical evidence in support of the technology acceptance model (TAM). Reviewing Sharma et al.'s (2009) method-method pair technique, Bagozzi (2011, p. 288) commented that the technique can "provide guidance on what methods or measures to accentuate or avoid in the future and what is to be learned by abstracting up from individual studies and looking for useful patterns of findings and conclusions, both methodologically and substantively".

Sharma et al.'s (2009) development of the method-method pair technique focused on estimating the magnitude of bias arising from one specific source of bias, viz. that due to the manner in which responses to survey items are captured. This paper extends Sharma et al.'s (2009) technique to estimate the magnitude of bias arising from another potential source of method bias, the manner in which the items for the predictor and criterion variables are worded. Podsakoff et al. (2012) argue that the content of items and similarities in item structure or wording can be sources of method bias. This could induce a correlated error in respondent scores for the two items, thus introducing a spurious component in the observed correlations (Sharma et al. 2009). Following from this, we hypothesize that the greater the similarity in the texts of the items employed to measure the predictor and criterion variables, the greater will be the magnitude of the spurious correlation.

This paper begins with a review of the literature on how the wording of survey items can induce a spurious correlation in observed correlations. It then develops the hypotheses to be tested in this research. This is followed by a review of natural language processing techniques for evaluating similarity of short text segment pairs, the technique employed in this study to test the proposed hypothesis. We then describe the methods employed to test the hypothesis. Specifically, we follow Sharma et al. (2009) and employ meta-analysis to test the hypothesis. The sample will then be described and the protocol employed for data collection outlined. We conclude with a discussion of the technique developed here, the implications for theory and practice, and our plans for extending this research.

Wording of survey items and method bias: a review

Survey methods typically involve presenting respondents with questions that provide textual stimuli and capturing their responses. Survey questions are designed to capture responses on measurable items that

operationalize a higher-order abstract construct of interest. There is a semantic relationship between items and the construct they are chosen to represent (Abbott 1997). An abstract construct may be defined as being comprised of multiple facets. Items employed to measure multiple facets of a single construct domain may include some overlap (Jarvis et al, 2003). This implicitly acknowledges the possibility that a single measurement item can be interpreted by a respondent to represent multiple domains, i.e. there is no one-to-one correspondence between the wording of an item and the construct it could evoke in a respondent's mind. Employing multiple items increases the chance that construct scores estimated from summing item scores do represent the construct that the items are supposed to measure.

While there are well-defined principles and practices surrounding the creation of reliable questionnaire items, research suggests that numerous factors can introduce errors and biases into measurement and, ultimately, produce biased estimates of construct-level relationships (MacKenzie et al. 2011; Podsakoff et al. 2012). The choice of question wording is also a potential source of bias (Podsakoff et al. 2012). Even slight variations in terminology between questions, such as "poverty" versus "welfare" (Smith 1987), or directionality of phrasing, such as negative or positive wording (Harvey et al. 1985), have been shown to result in markedly different responses. While the literature proposes a number of remedies to reduce ambiguity (Burton-Jones 2009), in the absence of a technique to estimate the magnitude of method bias on account of item wording, it is unclear if the proposed remedies actually mitigate the problem.

Respondents' interpretations of measurement items can be an important source of method bias. Respondents utilize a number of cognitive processes to interpret the meaning of a measurement item. These include identification of words, appropriate vocabulary and construction of meaning (Schwarz 1999). Often measurement items can be interpreted by respondents ambiguously. Ambiguity can exist for a number of reasons including issues of semantics, culture and context (Abbott 1997; Warnecke et al. 1997). Regardless of the source, ambiguity of interpretation is an outcome of cognitive processes involved as respondents "construct their own idiosyncratic meanings for them [items]" (Podsakoff et al. 2012, p. 551).

In most instances, when operationalizing a construct, it is often recommended to employ items that use similar words (Bagozzi, 2011, MacKenzie et al. 2011). However, employing similar words to operationalize different constructs would be undesirable. It would be especially problematic when similar words are employed in items measuring the predictor as well as the criterion variable. If a pair of text segments (i.e. items) measuring the predictor and criterion variables are similar, there is the possibility of a spurious correlation being induced between the measures of the two constructs. Clearly, this would be undesirable as the observed correlation would not accurately reflect the true correlation between the constructs (Sharma, 2009).

Extending the above arguments, we propose here that items representing predictor and criterion variables are interpreted idiosyncratically by respondents as representing multiple domains of meaning. Further, if the items share words and phrases, some of the idiosyncratic domains constructed by respondents for the predictor and criterion items could overlap. In which case, respondent scores on the predictor and criterion items will share some covariance on account of respondents' cognitive processes involved in interpreting the meaning of measurement items. Formally, we propose that

> H1: The greater the level of similarity between the items employed to measure the predictor variable and the items employed to measure the criterion variable, the greater will be the observed effect size between the predictor and criterion variables.

Natural language processing based techniques for measuring similarity of short text segment pairs

Testing the above hypothesis requires that researchers be able to rate the level of similarity of short text segment pairs, in this case a short text pair representing an item for a predictor variable and an item for a criterion variable. Here, we draw on natural language processing (NLP)-based techniques to rate short text pairs on similarity. NLP-based techniques have previously been extensively employed in IS research for similar applications. For instance, Sugumaran and colleagues apply NLP-based techniques to retrieve semantic-based information and improve the results of search queries on the web (Mëtais et al. 2013; Storey et al. 2008; Sugumaran and Storey 2003).

Humans utilise a number of cognitive processes to evaluate the extent to which two text segments are similar. Such heuristic judgments of similarity are employed by humans in a number of applications, such as organizing a large body of text into key themes, in writing a précis, and in summarizing a large body of text (Dale et al. 2000). NLP-based techniques attempt to identify protocols implicitly employed by humans in performing language processing tasks and then develop algorithms to mimic those protocols.

A number of NLP-based techniques have been proposed in the literature to automate language processing tasks by mimicking the manner in which humans perform those tasks. Such tasks include spelling correction (Budanitsky and Hirst 2001), text summarization (Dolan et al. 2004), machine translation (Wu and Palmer 1994), semantic based retrieval (Sugumaran and Storey 2003), and search. The techniques employed to automate these tasks essentially try to mimic human judgments employed in performing language processing tasks. Computational measures of similarity are a key component of many automated language processing tasks (Dolan et al. 2004; Manning et al. 2008; Mihalcea et al. 2006).

A number of similarity measures have been proposed in the literature for use in automated language processing tasks. Three broad classes of similarity measures are discussed in the literature: lexical-based measures, knowledge-based measures, and corpus-based measures (Meng et al. 2013; Mëtais et al. 2013; Metzler et al. ; Mihalcea et al. 2006; Rohde et al. 2009; Storey et al. 2008; Sugumaran and Storey 2003). Each measure draws upon a different set of heuristics employed in human judgments of similarity.
Lexical-based measures draw on the heuristic that human judgments of similarity are conditioned by the extent to which text segments share common words. For instance, the Jaccard similarity coefficient is based on the intersection of two text segments, i.e. the number of words shared between two text segments. Critics argue that human judgments of similarity are not based solely on exact words shared between text segments, but also take into account the overall meaning conveyed by the text segments to a rater (Dale et al. 2000; Li et al. 2006; Mihalcea et al. 2006).

Knowledge-based measures of similarity attempt to address the above limitation of lexical-based measures by trying to account for the meaning shared by individual words. To achieve this, those measures rely on a searchable knowledge base that maps the similarities and shared meanings of words. The knowledge-base is constructed by identifying multiple concepts commonly associated with a word, organizing them in a hierarchy and then connecting the concepts shared between words to create a searchable repository (Dale et al. 2000; Li et al. 2003). Similar to lexical-based measures, knowledge-based measures estimate similarity based on all words that are identical across text segments. In addition, they also account for words that are closely related in meaning, such as synonyms.

An important critique of knowledge-based measures is the availability of knowledge base repositories and their validities (Furlan et al. 2013; Rohde et al. 2009). Another critique is that human judgments of sentence pair similarity are not based solely on word-to-word meaning, but also take into account the context in which the words are used (Altmann and Steedman 1988).

Corpus-based measures of similarity are alternative approaches that attempt to address the limitations inherent in knowledge-based measures, such as availability and validity. These measures rely on calculating the number of times a particular pair of words appear together across a vast number of documents obtained from a large repository of text, such as the web. Corpus-based measures assume that the number of co-occurrences correlate positively with the word pair meaning. That is, the more frequently two words appear in close proximity in two texts, the higher the likelihood that the texts are similar. Consistent with lexical-based and knowledge-based measures, corpus based measures estimate similarity based on words that are identical across text segments, words that are closely related in meaning and words that co-occur in a large body of text, instead of in a limited knowledge base repository (Mihalcea et al. 2006; Rohde et al. 2009).

An important critique of corpus-based measures is that relying on a larger body of text does not improve the performance of a similarity measure against human judgment; instead, it only increases the computational complexity. The size of the repository is one factor among many that impacts the process of search and retrieval (Sugumaran and Storey 2003). Mihalcea et al. (2006) tested the performance of two

corpus-based and six different knowledge-based similarity measures on the Microsoft Research Paraphrase Corpus (MSRPC) data set and found that the performance of all measures was very similar.

The performance of similarity measures is typically evaluated by comparing the performance of the measures against standard datasets that have already been rated by human raters. For instance, the MSRPC has been frequently employed to evaluate the performance of similarity measures against human judgment (Achananuparp et al. 2008; Dolan et al. 2004; Meng et al. 2013). The MSRPC consists of 5801 sentence pairs that have been rated by human raters as 'Similar' or 'Not Similar'. Multiple human raters were employed in rating the sentence pairs in the MSRPC dataset and the consensus ratings of the human raters are provided as part of the dataset. Many studies reported in the literature have employed the MSRPC to evaluate the performance of different similarity measures. For instance, Mihalcea et al. (2006) and Achananuparp et al. (2008) tested the performance of a number of lexical-based, knowledge-based and corpus-based similarity measures on the MSRPC data set. Both studies report that the performance of all measures was very close to each other.

Prior studies evaluating the performance of lexical-based measures against knowledge-based or corpus-based measures suggest that while the performance of lexical based measures is similar to or better than other measures, they do not consistently outperform other measures. For instance, Metzler et al. (2007) and Achananuparp et al. (2008) found that lexical measures are good at finding semantically identical matches. This suggests that the lexical measures are still robust measures of human judgment of similarity (Dekai, 2010). This may also suggest that human judgments of similarity of two text segments rely strongly on the extent of lexical overlap.

Given the simplicity of lexical-based measures and their robust performance in previous research against the more sophisticated knowledge- and corpus-based measures, it was decided to employ a lexical-based measure of the similarity of predictor and criterion variable items to test H1. The Jaccard similarity coefficient is one of the most commonly employed measures of similarity and has performed well in prior research (Metzler et al. 2007 ; Achananuparp et al. 2008).

Method

Estimating the Jaccard similarity measure

The Jaccard similarity coefficient for two short text segments is defined as the function of the size of the intersection of two short sentences divided by the size of their union. The score is between 0 and 1, with 0 representing no similarity and 1 representing perfect similarity (i.e., almost a tautology).

As similar common words will inflate the similarity score, it is conventional to remove stop words in the text segments before calculating the Jaccard score and, indeed, other measures of similarity. Stop words are commonly occurring words, such as "a", "been", "so", "of" and "am" that do not have a significant influence on human interpretations of meaning. Consequently, they are removed from the computation of similarity scores. Including them in the computation of similarity scores could affect the similarity scores in a manner inconsistent with human judgments of similarity.

The other procedure employed after removing stop words and before the computation of the Jaccard score is stemming. Stemming refers to normalising text by transforming a word through stripping off its affixes without affecting the word context. Porter stemmer (http://tartarus.org/martin/PorterStemmer/), one of the most popular techniques for stemming English, is used in this research.

The Jaccard similarity score for two constructs is computed by extending the protocol for computing the score for a text segment pair. Operationalization of the predictor and criterion constructs are most likely to be accomplished by employing multi-item instruments for both constructs, with correlations reported between variables at construct level. Thus, text similarity is captured at item level then aggregated to the construct level. A matrix is created pairing each item measuring the predictor construct against each item measuring the criterion construct. For example, if a predictor (e.g. perceived usefulness) is operationalized by a 3 item instrument, and the criterion (e.g. use) is operationalized by a 4 item instrument, this will result in 3 x 4 = 12 predictor item-criterion item pairings. The Jaccard score for each predictor item-criterion item pairing is computed as per the above protocol for computing the Jaccard

score for two text segments. The Jaccard similarity score for two constructs is computed as the average of the Jaccard similarity scores of all possible predictor item-criterion item pairings for the two constructs.

A Python-based program was developed to take a file of text pairs (the predictor item text and the criterion item text), remove stop words, conduct a stemming procedure on the texts, compute the Jaccard similarity score for each text pair, and compute the study-level Jaccard similarity score. The Jaccard similarity scores generated by the program developed for this study were validated against the findings of Achananuparp et al. (2008). Following the protocol employed by Achananuparp et al. (2008), the program developed for this study was tested using MSRPC dataset. These findings suggest a high degree of reliability for the Jaccard similarity scores generated by the protocol employed in this research.

Analysis and hypothesis testing

The techniques of meta-analysis employed by Sharma et al. (2009) will be adapted to test the hypothesis developed here (H1). Sharma et al. followed the protocols of meta-analysis (Borenstein et al. 2009; Hunter and Schmidt 2004) to identify studies for inclusion in their meta-analysis. They then extracted the correlations, sample sizes and the method employed in the primary studies to measure the criterion variable from the included studies. The method employed to measure the criterion variable was rated on a theoretically derived scale of susceptibility to method bias. The hypothesis was then tested by meta-regressing (Borenstein et al. 2009) the predictor-criterion correlation reported in a study against the susceptibility of the study to method bias.

Following Sharma et al. (2009), H1 will be tested by meta-regressing the predictor-criterion correlations reported in the primary studies against their respective Jaccard similarity scores. Random-effects meta-regression is considered the most appropriate technique as it allows for the true effect size to vary across studies. As this meta-analysis reflects a random sample of studies capturing perceived use (PU) and use (U) in the TAM literature (see sample and data collection sections below), the effect sizes captured can also be assumed to reflect a random sample of effect sizes (Borenstein et al 2009). The statistical software program, Comprehensive Meta-Analysis-2 (CMA) will be used to perform the meta-regression analysis. General statistical packages, such as SAS and SPSS do not directly support random-effects meta-regression. The parameters estimated by standard random-effects regression protocols in general statistical packages cannot be directly interpreted according to the parameters of a meta-regression analysis (Lipsey & Wilson 2001). A number of plug-ins are available to adapt the standard protocols in general statistical packages for performing random-effects meta-regression (Lipsey and Wilson 2001). However, with the availability of dedicated meta-analysis packages, such as CMA, it is preferable to employ a dedicated package. A random effects meta-regression will be employed to test H1 on two independent samples (see below) according to the protocol suggested by Borenstein et al. (2009).

In Sample 1, the criterion variable for the meta-regression will be the correlation between perceived usefulness (PU) and use (U) and the predictor variable will be the study-level Jaccard similarity score. In Sample 2, the criterion variable for the meta-regression will be the correlation between attitude and physical activity behavior and the study-level Jaccard similarity score will be the predictor variable.

Computing the Jaccard similarity score for a study requires that the study report the text for the items employed to operationalize the predictor and the criterion constructs. Studies included in this meta-analysis (see Sample below) have been examined and item text extracted. In many cases, complete texts of the items were not reported. If sufficient information was reported, item texts were reconstructed; in other cases, if references to standard instruments, (e.g. the Davis perceived usefulness instrument (Davis 1989), were provided, the item text was reconstructed. Where item text was not reported or could not be reconstructed, studies were not included in this meta-analysis.

Sample and data collection

The hypothesis (H1) developed above will be tested on two samples of data. Examining H1 on two distinct samples allows us to draw conclusions about the generalizability of the technique developed here to explain the magnitude of method bias across studies and to suggest directions for future research to further develop the technique. The first sample is the set of studies included in Sharma et al.'s (2009) meta-analysis of the TAM literature. The 76 studies included in Sharma et al.'s (2009) meta-analysis were examined to extract the text of the items employed to operationalize perceived usefulness (PU) and use

(U). Where the item texts were not reported or could not be reliably reconstructed, the study could not be included in this meta-analysis. Our protocol identified 50 studies for inclusion in this meta-analysis.

The second sample was chosen to be contextually distinct from the first sample. The two samples correspond in that the TAM PU-U correlation is consistent with the attitude-behavior correlation from the Theory of Reasoned Action/Theory of Planned Behavior (TRA/TPB) model (Fishbein and Ajzen 1974; Fishbein and Ajzen 1975). However, the two samples differ in terms of the behavior under investigation; the TAM studies in Sample 1 examine technology use behavior while Sample 2 comprises studies investigating physical activity behavior.

The sample for the second dataset was collected employing standard search protocols for locating studies for inclusion in a meta-analysis (Borenstein et al. 2009; Sharma et al. 2004; Sharma et al. 2009). A systematic literature search was employed to identify studies that investigated the attitude-behavior relationship in the context of physical activity behavior. Bibliographic data bases such as Scopus, Medline, PsychInfo, Google Scholar, and Dissertation Abstracts were searched using the search terms attitude, physical activity, exercise, theory of planned behavior, theory of reasoned action, health behavior, and behavior change. Electronic versions of key journals publishing in the area of physical activity were also searched for relevant studies. These included the International Journal of Behavioral Nutrition and Physical Activity, Medicine and Science in Sports and Exercise, and the Journal of Sport and Exercise Psychology. The search period was limited to studies published between 2000 and 2013. Additionally, studies included in meta-analyses of the theory of planned behaviour and physical activity by Chatzisarantis (2011), Hagger (2002), Hagger and Chatzisarantis (2009), and McEachen (2011) were examined.

Following Borenstein et al. (2009) and Hunter and Schmidt (2004), two criteria were employed to include studies in this meta-analysis. First, the study should report the correlation (or equivalent effect size metric) between attitude and physical activity, as well as the sample size. Second, the study should report in detail the items employed to measure attitude and physical activity. This protocol identified 131 data points from 76 publications (60 journal articles and 16 dissertations) for Sample 2.

Results

The two datasets are currently being analyzed and results will be available at the time of ICIS 2014.

Discussion and directions for future research

This paper has outlined a program of research to develop techniques to estimate the magnitude of method bias induced in observed effect sizes as a result of the level of similarity in the texts of items employed to measure the predictor and criterion variables. These techniques have important implications for both research and practice.

For researchers, these techniques provide an easily operationalizable and valid technique to partial out the effects of method bias due to text similarity. This enables robust testing of theory by enabling researchers to partial out the effect of method bias. This technique complements and extends Sharma et al.'s (2009) method-method pair technique by developing a technique to estimate bias due to an additional source. Sharma et al.'s (2009) method-method pair technique enables the estimation of the magnitude of method bias arising from the manner in which responses to survey items are collected. Extending the method-method pair technique, the item similarity technique developed here would enable the estimation of the magnitude of method bias arising from the wording of measurement items. As outlined in Sharma et al. (2009), the intercept of the meta-regression is an estimate of the correlation between the constructs, controlling for the effect of various sources of method bias. Following Bagozzi (2011, p. 288), the item similarity technique also enables researchers to empirically evaluate good practices for the design of survey instruments by evaluating "what ... measures to accentuate or avoid in the future".

For practitioners, this technique has application and relevance as a tool for designing robust survey instruments that minimize the extent of method bias in the data captured. Practitioners commonly employ survey-based research in various settings, such as public policy development, clinical research, business decision-making, and in market research. This technique would enable practitioners to a priori

predict the magnitude of spurious correlation that can be expected for specific relationships. As outlined in Sharma et al. (2009), the intercept of the meta-regression is an estimate of the correlation between the constructs, controlling for the effect of various sources of method bias. This can enable both the improvement of survey design as well as more valid conclusions that could be drawn from survey data.

While promising, the development of this technique requires additional testing. First, the protocol developed here to estimate lexical similarity is in need of further validation. Currently, the performance of this technique against other NLP-based techniques is encouraging. However, the performance of the technique against human judgment in survey settings needs to be investigated. A Q-sort protocol for testing the technique against human judgment is under development. Second, the technique will be extended to evaluate the performance of knowledge- and corpus-based measures of similarity against human judgment.

Further, the NLP literature itself is refining similarity measures to make them better representations of human judgment. As an example, consider the following three text segments; 1) I like apples very much; 2) I like oranges very much; 3) I just love apples. (Thank you to the Associate Editor for this example.) Broadly speaking, the Jaccard lexical similarity measure calculates the ratio of the number of identical words (the intersection set) to the total number of unique words across the two text segments (the union set). As text segments 1 and 2 share 4 words ("I", "like", "very" and "much") of a total of 6 unique words ("I", "like", "very", "much", "apples", and "oranges"), the similarity score is 4/6. For text segments 1 and 3, the similarity score is 2/7, while that for text segments 2 and 3 it is 1/8.

In the above example, a human rater will likely judge that text segments 1 and 3 are more similar than text segments 1 and 2. However, the Jaccard similarity measure would suggest the opposite conclusion. While it is the objective of similarity measures to mimic human judgment, there remains room to improve on those measures. As a direction for future research, we will be employing knowledge-based as well as corpus-based measures to estimate the similarity of item pairs.

Conclusion

This paper has proposed a technique and a program of research to estimate the magnitude of bias arising from the wording of items employed in surveys to capture the predictor and criterion variables. The technique is based on natural language processing-based measures of text similarity. The technique will be employed to test the hypothesis that the greater the level of similarity between items employed to measure the predictor and criterion variables, the greater the observed effect size between those variables. Two samples will be employed to examine this technique and results will be reported at ICIS 2014.

Acknowledgements

The support provided for this research by the Australian Research Council Discovery grant (DP130100068) is gratefully acknowledged. The authors extend their thanks to Andrew Burton-Jones and Pascal Perez for their invaluable comments and input on earlier versions of this paper. The authors also gratefully acknowledge the dedicated work of Madalyn Oliver in data collection and consolidation.

References

Abbott, A. 1997. "Seven Types of Ambiguity," *Theory and Society* (26:2/3), pp. 357-391.

Achananuparp, P., Hu, X., and Shen, X. 2008. "The Evaluation of Sentence Similarity Measures," in *Data Warehousing and Knowledge Discovery*. Springer, pp. 305-316.

Altmann, G., and Steedman, M. 1988. "Interaction with Context During Human Sentence Processing," *Cognition* (30:3), pp. 191-238.

Avolio, B.J., Yammarino, F.J., and Bass, B.M. 1991. "Identifying Common Methods Variance with Data Collected from a Single Source: An Unresolved Sticky Issue," *Journal of Management* (17:3), pp. 571-587.

Bagozzi, R. 2011. "Measurement and Meaning in Information Systems and Organizational Research: Methodological and Philosophical Foundations," *MIS Quarterly* (35:2), pp. 261-292.

Borenstein, M., Hedges, L.V., Higgins, J.P.T., and Rothstein, H.R. 2009. *Introduction to Meta-Analysis*. West Sussex, UK: Wiley.

Budanitsky, A., and Hirst, G. 2001. "Semantic Distance in Wordnet: An Experimental, Application-Oriented Evaluation of Five Measures," *Workshop on WordNet and Other Lexical Resources*.

Burton-Jones, A. 2009. "Minimizing Method Bias through Programmatic Research," *MIS Quarterly* (33:3), pp. 445-471.

Campbell, D.T., and Fiske, D.W. 1959. "Convergent and Discriminant Validation by the Multitratit-Multimethod Matrix," *Psychological Bulletin* (56:2), pp. 81-105.

Chin, W.W., Thatcher, J.B., and Wright, R.T. 2012. "Assessing Common Method Bias: Problems with the Ulmc Technique," *MIS Quarterly* (36).

Cook, T.D., and Campbell, D.T. 1979. *Quasi-Experimentation: Design and Analysis Issues for Field Studies*. Chicago: Rand McNally.

Dale, R., Moisl, H., and Somers, H.L. 2000. *Handbook of Natural Language Processing*. New York: Marcel Dekker.

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

Dekai, W. 2010. "Alignment," in *Handbook of Natural Language Processing,* N. Indurkhya and F.J. Damerau (eds.). CRC Press, pp. 367-408.

Dolan, B., Quirk, C., and Brockett, C. 2004. "Unsupervised Construction of Large Paraphrase Corpora: Exploiting Massively Parallel News Sources," *Proceedings of the 20th International Conference on Computational Linguistics, 2004*.

Fishbein, M., and Ajzen, I. 1974. "Attitudes toward Objects as Predictors of Single and Multiple Behavioral Criteria," *Psychology Review* (81:1), pp. 59-74.

Fishbein, M., and Ajzen, I. 1975. *Belief, Attitude, Intention and Behavior: An Introduction to Theory and Research*. Reading, MA: Addison-Wesley.

Furlan, B., Batanović, V., and Nikolić, B. 2013. "Semantic Similarity of Short Texts in Languages with a Deficient Natural Language Processing Support," *Decision Support Systems* (55:3), pp. 710-719.

Harvey, R.J., Billings, R.S., and Nilan, K.J. 1985. "Confirmatory Factor Analysis of the Job Diagnostic Survey: Good News and Bad News," *Journal of Applied Psychology* (70:3), pp. 461-468.

Hunter, J.E., and Schmidt, F. 2004. *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings*, (2nd ed.). Newbury Park, CA: Sage.

Jarvis, C.B., MacKenzie, S.B., and Podsakoff, P.M. 2003. "A Critical Review of Construct Indicatorsand Measurement Model Misspecification in Marketing and Consumer Research," *Journal of Consumer Research* (30:2), pp. 199-218.

Li, Y., Bandar, Z.A., and McLean, D. 2003. "An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources," *Knowledge and Data Engineering, IEEE Transactions on* (15:4), pp. 871-882.

Li, Y., McLean, D., Bandar, Z.A., O'shea, J.D., and Crockett, K. 2006. "Sentence Similarity Based on Semantic Nets and Corpus Statistics," *Knowledge and Data Engineering, IEEE Transactions on* (18:8), pp. 1138-1150.

Lindell, M.K., and Whitney, D.J. 2001. "Accounting for Common Method Variance in Cross-Sectional Research Designs," *Journal of Applied Psychology* (86:1), pp. 114-121.

Lipsey, M.W., and Wilson, D.B. 2001. *Practical Meta-Analysis*. Thousand Oaks: Sage Publications.

MacKenzie, S.B., Podsakoff, P.M., and Podsakoff, N.P. 2011. "Construct Measurement and Validity Assessment in Behavioral Research: Integrating New and Existing Techniques," *MIS Quarterly* (35:2), pp. 293-334.

Manning, C.D., Raghavan, P., and Schütze, H. 2008. *Introduction to Information Retrieval*. New York: Cambridge University Press.

Meng, L., Huang, R., and Gu, J. 2013. "A Review of Semantic Similarity Measures in Wordnet," *International Journal of Hybrid Information Technology* (6:1).

Mëtais, E., Meziane, F., Saraee, M., Sugumaran, V., Vadera, S., SpringerLink, and Springer, V. 2013. *Natural Language Processing and Information Systems: 18th International Conference on Applications of Natural Language to Information Systems, Nldb 2013, Salford, Uk, June 19-21, 2013. Proceedings*. Berlin, Heidelberg: Springer Berlin Heidelberg.

Metzler, D., Dumais, S., and Meek, C. 2007 "Similarity Measures for Short Segments of Text," pp. 16-27.

Mihalcea, R., Corley, C., and Strapparava, C. 2006. "Corpus-Based and Knowledge-Based Measures of Text Semantic Similarity." pp. 775-780.

Podsakoff, P.M., MacKenzie, S.B., Lee, J.-Y., and Podsakoff, N.P. 2003. "Common Method Biases in Behavioral Research: A Critical Review of the Literature and Recommended Remedies," *Journal of Applied Psychology* (88:5), pp. 879-903.

Podsakoff, P.M., MacKenzie, S.B., and Podsakoff, N.P. 2012. "Sources of Method Bias in Social Science Research and Recommendations on How to Control It," *Annual Review of Psychology* (65), pp. 539-569.

Richardson, H.A., Simmering, M.J., and Sturman, M.C. 2009. "A Tale of Three Perspectives: Examining Post Hoc Statistical Techniques for Detection and Correction of Common Method Variance," *Organizational Research Methods* (12:4), pp. 762-800.

Rohde, D.L., Gonnerman, L.M., and Plaut, D.C. 2009. "An Improved Model of Semantic Similarity Based on Lexical Co-Occurrence," *Cognitive Science*), pp. 1-33.

Schwarz, N. 1999. "Self-Reports: How the Questions Shape the Answers," *American Psychologist* (54:2), pp. 93-105.

Sharma, R., Yetton, P., and Crawford, J. 2004. "The Relationship between Perceived Usefulness and Use: Estimating Common Methods Bias," *Meeting of Special Interest Group on Adoption and Diffusion of Technologies*, Washington, DC.

Sharma, R., Yetton, P., and Crawford, J. 2010. "A Critique of the Marker Variable Technique: The Effect of Alternative Marker Variable Criteria," *18th European Conference on Information Systems*, Pretoria, South Africa.

Sharma, R., Yetton, P.W., and Crawford, J. 2009. "Estimating the Effect of Common Method Variance: The Method-Method Pair Technique with an Illustration from Tam Research," *MIS Quarterly* (33:3), pp. 473-490.

Smith, T.W. 1987. "That Which We Call Welfare by Any Other Name Would Smell Sweeter an Analysis of the Impact of Question Wording on Response Patterns," *The Public Opinion Quarterly* (51:1), pp. 75-83.

Storey, V.C., Burton-Jones, A., Sugumaran, V., and Purao, S. 2008. "Conquer: A Methodology for Context-Aware Query Processing on the World Wide Web," *Information Systems Research* (19:1), pp. 3-25.

Straub, D.W.J., and Burton-Jones, A. 2007. "Veni, Vidi, Vici: Breaking the Tam Logjam," *Journal of the Association for Information Systems* (8:4), pp. 223-229.

Sugumaran, V., and Storey, V.C. 2003. "A Semantic-Based Approach to Component Retrieval," *Data Base for Advances in Information Systems* (34:3), pp. 8-24.

Warnecke, R.B., Johnson, T.P., Chávez, N., Sudman, S., O'rourke, D.P., Lacey, L., and Horm, J. 1997. "Improving Question Wording in Surveys of Culturally Diverse Populations," *Annals of epidemiology* (7:5), pp. 334-342.

Wu, Z., and Palmer, M. 1994. "Verb Semantics and Lexical Selection," *Proceedings of the 32nd annual meeting of Association for Computational Linguistics*. Association for Computational Linguistics, pp. 133-138.