Journal of Informetrics 12 (2018) 158–180

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

A scalable and adaptive method for finding semantically equivalent cue words of uncertainty

Chaomei Chen[a,b], Min Song[b,*], Go Eun Heo[b]

[a] College of Computing and Informatics, Drexel University, USA
[b] Department of Library and Information Science, Yonsei University, South Korea

ARTICLE INFO

Article history:
Received 16 May 2017
Received in revised form
10 December 2017
Accepted 10 December 2017

Keywords:
Uncertainty
Semantically equivalent words
Scientific assertions
Deep learning
Resources

ABSTRACT

Scientific knowledge is constantly subject to a variety of changes due to new discoveries, alternative interpretations, and fresh perspectives. Understanding uncertainties associated with various stages of scientific inquiries is an integral part of scientists' domain expertise and it serves as the core of their meta-knowledge of science. Despite the growing interest in areas such as computational linguistics, systematically characterizing and tracking the epistemic status of scientific claims and their evolution in scientific disciplines remains a challenge. We present a unifying framework for the study of uncertainties explicitly and implicitly conveyed in scientific publications. The framework aims to accommodate a wide range of uncertainty types, from speculations to inconsistencies and controversies. We introduce a scalable and adaptive method to recognize semantically equivalent cues of uncertainty across different fields of research and accommodate individual analysts' unique perspectives. We demonstrate how the new method can be used to expand a small seed list of uncertainty cue words and how the validity of the expanded candidate cue words is verified. We visualize the mixture of the original and expanded uncertainty cue words to reveal the diversity of expressions of uncertainty. These cue words offer a novel resource for the study of uncertainty in scientific assertions.

1. Introduction

A scientific proposition is a statement such as smoking causes cancer. The epistemic status of a scientific proposition refers to the best knowledge of its truthfulness given the current scientific knowledge. Thus, the epistemic status may range from completely unknown to speculations and from hypotheses to facts. The concept of uncertainty in this context characterizes the lack of sufficient information on a given proposition. A statement concerning a proposition can be considered as a combination of two parts: the proposition proper and information relevant to the epistemic status of the proposition. In this article, we focus on uncertainties due to lack of information and, in particular, uncertainties due to lack of consensus.

Scientists routinely deal with such uncertainties at various stages of their research, from formulating research questions and selecting research methods to interpreting their findings and communicating their work to others (Cordner & Brown, 2013). Light, Qiu, & Srinivasan (2004) estimated that 11% of sentences in MEDLINE abstracts are speculative. Sociologists have studied the formation of consensus in the scientific community concerning whether smoking indeed causes cancer and whether a consensus is reached on climate change (Shwed & Bearman, 2010). Scientists face intensified uncertainties when inconsistent, conflicting, or contradictory findings emerge and when competing paradigms are proposed to resolve

* Corresponding author.
E-mail address: min.song@yonsei.ac.kr (M. Song).

https://doi.org/10.1016/j.joi.2017.12.004
1751-1577/

pressing crises (Kuhn, 1970). The formation of a consensus or the establishment of a dominant paradigm may correspond to a decrease of the overall uncertainty associated with a field of research. However, as we all know, searching for answers to seemingly simple questions may quickly lead to many complicated questions. The ability to assess the state of the art of a field of research effectively and efficiently at various levels of granularity is crucial for scientists, science policy makers, and the public.

Research in computational linguistics has made significant advances in identifying uncertainty cues and negations. Remarkably influential efforts include the development of the BioScope Corpus for uncertainty and negation in biomedical publications (Vincze, Szarvas, Farkas, Móra, & Csirik, 2008), the CoNLL 2010 Shared Task (Farkas, Vincze, Móra, Csirik, & Szarvas, 2010) for detecting hedges and their scope in natural language texts, the enrichment of a biomedical event corpus with meta-knowledge (Thompson, Nawaz., McNaught, & Ananiadou, 2011), and unifying categorizations of semantic uncertainty for cross-genre and cross domain uncertainty detection (Szarvas et al., 2012).

For example, the CoNLL-2010 shared task (Farkas et al., 2010) focused on detection of uncertainty cues and its linguistic scope in natural language texts. A typical hedging cue is composed of four categories: 1) auxiliaries, 2) verbs of hedging or verbs with speculative content, 3) adjectives or adverbs, and 4) conjunctions. Up to now, uncertainty detection has focused on biomedical articles and text on Wikipedia. According to Farkas et al. (2010), the best uncertainty detection performance in the CoNLL-2010 shared task was achieved with sequence labeling (e.g., Conditional Random Fields) in the biomedical data and bag of words sentence classification in the Wikipedia data. For the in-sentence hedge scope detection task, they classify each token to detect specific cue scopes. More recent studies have explored the potential of measuring the confidence of biomedical models such as pathways based on textual uncertainty (Zerva, Batista-Navarro, Day, & Ananiadou, 2017) and the feasibility of assessing the factuality of semantic predications (Kilicoglu, Rosemblat, & Rindflesch, 2017). Kilicoglu et al. (2017) define factuality as a degree of uncertainty that has seven values, namely fact, probable, possible, doubtful, counterfact, uncommitted, and conditional.

In a broader context, identifying and measuring the degree of uncertainties associated with scientific knowledge embedded in the vast and fast-growing volume of scientific literature remain a bottleneck (Chen, 2016). Influential computational linguistic approaches such as hedging (Hyland, 1998), semantic uncertainty (Szarvas et al., 2012), negation (Chapman, Bridewell, Hanbury, Cooper, & Buchanan, 2001; Morante & Daelemans, 2009), and discourse-level uncertainty (Vincze, 2013) have been largely motivated by issues concerning uncertainties from linguistic perspectives. As demonstrated by Simmerling and Janich (2015), by using grammatical, stylistic, and rhetorical options, one can talk about scientific uncertainty without using any lexical cues of uncertainty. Furthermore, philosophical and sociological studies of science, scientific creativity, and scientific discovery have highlighted the role of identifying and resolving contradictions and inconsistencies in scientific discovery and in divergent thinking in general. In particular, the value of reconciling multiple perspectives has been long recognized and advocated (Collins, 1989; Linstone, 1981). It is critical for scientists to be able to track conflicting views on the same issue and resolve seemingly contradictory evidence at a new level (Chen, 2014, 2016). The linguistically motivated approaches to the study of scientific uncertainty may benefit from a broadened scope of perspectives.

In this article, we present a conceptual framework of the study of uncertainty based on a novel conceptualization of uncertainty as an epistemic status of scientific propositions. The new conceptualization underlines the nature of uncertainty as a meta-knowledge of science and its integral role in scientific change. We introduce a scalable and adaptive method to identify uncertainty cues under the broadened conceptualization of uncertainty. The resultant uncertainty cue words are expected to provide a useful resource for further studies of scientific uncertainty. The method is adaptive in the sense that analysts may generate semantically equivalent uncertainty cues of new dimensions based on a small number of example words.

The rest of the article is organized as follows. First, we introduce basic concepts concerning scientific propositions and illustrate some of the most common types of uncertainties associated semantic predications in MEDLINE and the distributions of leading uncertainty cue words in other collections of scientific publications. Next, we present a scalable and adaptive method to construct a comprehensive set of uncertainty cue words from scientific publications. The method begins with a set of hand-crafted uncertainty cue words as seeds based on a general-purpose thesaurus of English. Then the computational method expands the seed list to a much larger set of semantically equivalent uncertainty cue words. Two judges evaluated the expanded cue words. The accepted and rejected cue words along with the seed words are visualized as non-overlapping clusters. Sample sentences selected by these uncertainty cues are discussed. The collection of the specific uncertainty cue words, classes of these words, and corresponding statistics are provided as a community resource for researchers to build on the result of our research.

2. Uncertainties of scientific knowledge

Scientific knowledge is a complex adaptive system of facts, beliefs, hypotheses, speculations, opinions, and a wide variety of other types of information about what we know and how much we know. It is adaptive in that existing scientific knowledge is subject to re-examination in light of new discoveries, alternative interpretations, and scenarios that are previously thought impossible (Chen, 2014; Popper, 1961). A scientist's domain expertise consists of not only his or her knowledge of various facts and consensus in science but also an accurate understanding of the epistemic status of a wide variety of unsettled elements of a scientific domain. The epistemic status of a scientific proposition characterizes various stages of its epistemological advances driven by underlying scientific inquiries. For example, our beliefs of the truthfulness of a proposition may vary