Epistemic modality: From uncertainty to certainty in the context of information seeking as interactions with texts

Victoria L. Rubin *

Faculty of Information and Media Studies, University of Western Ontario, London, Ontario, Canada N6A 5B7

Abstract

This article introduces a type of uncertainty that resides in textual information and requires epistemic interpretation on the information seeker’s part. Epistemic modality, as defined in linguistics and natural language processing, is a writer’s estimation of the validity of propositional content in texts. It is an evaluation of chances that a certain hypothetical state of affairs is true, e.g., definitely true or possibly true. This research shifts attention from the uncertainty–certainty dichotomy to a gradient epistemic continuum of absolute, high, moderate, low certainty, and uncertainty. An analysis of a New York Times dataset showed that epistemically modalized statements are pervasive in news discourse and they occur at a significantly higher rate in editorials than in news reports. Four independent annotators were able to recognize a gradation on the continuum but individual perceptions of the boundaries between levels were highly subjective. Stricter annotation instructions and longer coder training improved intercoder agreement results. This paper offers an interdisciplinary bridge between research in linguistics, natural language processing, and information seeking with potential benefits to design and implementation of information systems for situations where large amounts of textual information are screened manually on a regular basis, for instance, by professional intelligence or business analysts.

1. Introduction

The systems-centered approach defines information retrieval (IR) most generally as being concerned with representation, storage, organization, and accessing of information items (Salton & McGill, 1983). The user-centered approach re-conceptualizes IR as an information seeking activity supporting interactions with texts as the central process (Belkin, 1993). Information seeking in interactive IR is acquisition of information from knowledge sources (Ingwersen & Järvelin, 2004). Users are seen as not just passive recipients of messages but active seekers for texts of potential interests: “they [users] make judgments about the usefulness or interest of texts by engaging with them, they interpret texts in order to understand them” (Belkin, 1993; p. 55).

In interactions with IR and other information systems, information seekers encounter at least two types of uncertainties. The first type is a psychological phenomenon, both a cognitive and affective state, and an experience related to the task at hand (discussed in Section 1.1); and the second type is a linguistic and epistemic phenomenon in texts that captures the source’s estimation of a hypothetical state of affairs being true. The latter is associated with the content of the retrieved information and is defined in Section 1.2 as epistemic modality, and is central to this study.

Texts carry multiple shades of linguistic and epistemic uncertainties expressed from different sources’ perspectives. Writers often overtly position their qualifications of facts or opinions on an epistemic continuum ranging from uncertainty to...
absolute certainty. For instance, any statement or claim may be modalized by the writer's qualification about whether the statement is likely to be true, e.g., “Yeltsin, for one last time, seemed to be betting that an impulsive, bold move would turn Russia toward a more promising future.” New York Times, 2000. The writer's moderate certainty, in this case, is relevant to the information seeking (IS) context since users routinely assess and de-code such subtle linguistic cues as seemed to be betting in order to make judgments of the validity of the retrieved content. Whether they are aware of it or not, information seekers are often presented with these cues about attitudes towards what is being said and estimations of the likelihood that the content is true. The process ultimately contributes to the information seeker's subjective interpretation of the situation and the traditional IS uncertainty type of a cognitive, affective, and experiential nature.

Native speakers are usually unaware of the complexity of the language tasks they perform so effortlessly. Yet bringing this linguistic awareness into the IS context can potentially increase flexibility of the retrieval system designs, for instance, by providing pre-defined choices of the desired epistemic strength of the retrieved information.

In the remainder of the article, I will review the underlying assumptions regarding traditional uncertainty in the IS context as seen by several well-established models, and present the conceptual and analytical framework for linguistic certainty, or epistemic modality, originating from natural language computing literature (Rubin, 2006; Rubin, Kando, & Liddy, 2004; Rubin, Liddy, & Kando, 2005; Wiebe, 1990). The methods and results of the empirical text annotation and intercoder reliability study will follow. I will conclude with the implications of epistemic modality awareness in interaction with texts for information system design in the IS context.

1.1. The object and locus of uncertainty in information seeking

What is the object of uncertainty in IS? The concept of uncertainty in information seeking and use literature has been equated to an information need itself (Dervin, 1983), described as a cognitive gap (Yoon & Nilan, 1999), and raised to the basic principle of information seeking (Kuhlthau, 1993). What has not been clearly summarized, to the best of my knowledge, is what uncertainty may be about in the IS context.

Uncertainty is most often associated with the experience of making relevance judgments about the retrieved information and with its fit to the information seeker's need. Attfield and Dowell (2003) describe uncertainty as perceived relevance or potential usefulness of information. Yoon and Nilan (1999) also see both certainty and uncertainty as the user's perception aspects of a cognitive behavior. In Kuhlthau (1999) model of the information search and process, uncertainty is seen as an affective and cognitive experience with a varying intensity of uncertainty depending on the stage of IS: initiation, selection, exploration, formulation, collection, and presentation. Kuhlthau's uncertainty, confusions, and doubt that eventually give way to confidence are related to the process of information seekers focusing there IS task and constructing new understanding. Wilson, Ford, Ellis, and Foster (2002) simplify Kuhlthau (1999) model to four stages: problem recognitions, refining, and finding relevant information to assist them in resolving their problem (Wilson, 1999; Wilson et al., 2002).

In summary, the cognitive, affective, and experiential kind (or, arguably, kinds) of uncertainty is about the perceived information relevance, usefulness, availability, or potential success in focusing the task and resolving the information need (or problem). These are accepted, much discussed, and perfectly valid objects of uncertainty in IS. It is unclear, however, why the spectrum of what is important to the information seeker's experience does not extend to the evaluation of the content of the retrieved information. While the line between IS and information use is blurry (Anderson, 2006), information seekers inevitably have to interpret some of the retrieved information (be it unique or redundant to the existing knowledge, in Kuhlthau (1993) terms), in order to, at least, evaluate its relevance. Furthermore, virtually any retrieved proposition in texts can have an uncertainty associated with it.

Where does uncertainty originate in IS? Uncertainty in the IS literature seems to originate in the information seeker's consciousness, by the virtue of being a cognitive, or affective state, or an experience due to arising information need, or a lack of knowledge (Anderson, 2006; Attfield & Dowell, 2003; Dervin, 1983; Kuhlthau, 1999; Wilson, 1999; Wilson et al., 2002; Yoon & Nilan, 1999). To the best of my knowledge, it has not yet been articulated in the IS uncertainty research that information-bearing objects themselves are a source of uncertainty for information seekers. Information evaluation does not necessarily stop at the point of deciding if the retrieved information is relevant. Further transfer of ideas takes place. The retrieved information, even if it satisfies the information need, is not necessarily a coherent idea that exhibits ultimate confidence in the state of affairs and simply moves the information seeker from uncertainty to knowledge. Neither is it, under most circumstances, a straight-forward answer or a missing puzzle piece that will definitively fill in the user's cognitive gap. Most likely, the user will face a heterogeneous discourse with multiple statements with varying levels of certainty. Thus, there is a need for discussion about whether another kind of uncertainty may be inherently encoded in information itself, and be potentially transferred to the information seeker in the evaluation process.

An additional point worth emphasizing is that texts themselves are simply a conduit for ideas ultimately expressed by other humans. I entirely agree with Yoon and Nilan (1999) that the information seekers' interaction with texts should be considered as human-to-human communication in the IS context. In essence, no matter what the IS mediation means and environments are, be it IR systems or simple print newspapers, it is ultimately an exchange of knowledge and ideas between humans. Thus, the focus shifts from texts themselves to whose idea was expressed in the texts.
How is uncertainty related to certainty? Yoon and Nilan (1999) call to shift an emphasis from user's uncertainty to the inclusion of user's certainty. They focus on users’ perceptions of an information need and a linguistic articulation of a particular need situation (i.e., uncertainty) specified largely in terms of what users know (i.e., certainty). Anderson (2006) also notices certainty in her interviewees’ judgments about their search task at hand: "... So, clearly, it is something that I need to be familiar with..." (p. 14) and refers to a certainty–uncertainty spectrum. Even though Anderson talks about narrower scope of experiential certainty found in transcripts of IS interviewees, she observes linguistic cues such as “clearly, I don’t know, I’m not really confident, I’m not sure, it’s not clear to me” (p. 9), which are, in fact, epistemic modality markers. Anderson’s choice of the term “spectrum” is appropriate, even though it is not associated with the texts themselves, and is only illustrated with two traditionally recognized categories: certain or uncertain. My further argument is that these two states should not be viewed in a binary fashion but rather as a gradual epistemic continuum with several articulated points in between. Uncertainty is just one extreme on an epistemic continuum ranging from doubt and confusion to the highest attainable point of knowing, an implicit assurance, “an assumed truism, never called in question” in Wittgenstein’s terms (Moyal-Sharrock, 2004). The boundaries between shades of epistemic states are subtle and are rather subjectively interpreted from user to user. Before I set out to present evidence to this fact, I will re-introduce the concept of epistemic modality.

1.2. Epistemic modality, or certainty, in natural language processing

The field of natural language processing (NLP, or computational linguistics) studies the structure, function, and use of language, and organizes it into computational models for the design and development of a wide range of language-related software applications (Joshi, 1999). NLP undertakes the task of analyzing different kinds of information encoded in a free text form for the purpose of systems development; it is uniquely positioned to make systems improvement recommendations for user-centered IS and IR applications.

In its broadest sense, text is an ancient technology of information transfer that embodies most of human knowledge (Britton & Black, 1985). Text can convey more than just a writer’s propositional context of assertions (Coates, 1987), e.g., "A is true." Text can also transfer the writers’ attitudes to the propositions, assessments of possibilities, and the writer’s certainty, or lack thereof, in the validity of the truth of the statements, e.g., "it is possible that A is true or B is confident A is true." A statement is qualified in such a way (beyond its mere referential function) is modal, or epistemically modalized (Coates, 1987; Westney, 1986).

Epistemic modality, or certainty, concerns a linguistic expression of an estimation of the likelihood that a certain hypothetical state of affairs is, has been, or will be true (Nuyts, 2001). Subtle linguistic clues, or markers, contribute toward the user's understanding of how much credibility can be attached to individual propositions and whether the information comes from the first-hand or second-hand sources.

Pragmatics is a branch of linguistics that “seeks to explain the meaning of linguistic messages in terms of their context of use” (Mitkov, 2003, p. 137). “Discourse is an extended sequence of sentences produced by one or more people with the aim of conveying or exchanging information” (Mitkov, 2003, p. 113). Both pragmatic and discourse linguistic literatures are abundant in discussions of epistemic modality (Coates, 1987; Nuyts, 2001); mood (Palmer, 1986); evidentiality and evidentials (Mushin, 2001); expressions of doubt and certainty (Holmes, 1982; Hoye, 1997) and hedging (Hyland, 1998, 1999; Lackoff, 1972; Zuck & Zuck, 1986). Little attempt, however, has been made into manually annotate and consequently automate identification of statements with an explicitly expressed certainty or doubt, or shades of epistemic qualifications in between. This lack is possibly due to the complexity of computing epistemic interpretations in different pragmatic contexts; and due to unreliability of variety of linguistic expressions in English that could explicitly qualify a statement. Another complication is a lack of agreed-upon and easily identifiable discrete categories on the continuum from certainty to doubt. Several annotation projects have successfully addressed closely related subjective issues such as private states in news writing (Wiebe, Wilson, & Cardie, 2005) and hedging in scientific writing (Light, Qiu, & Srinivasan, 2004; Mercer, DiMarco, & Kroon, 2004). Having access to the opinion holder’s evaluation of how true a statement is valuable in predicting reliability of arguments and claims, and can potentially enhance information system design by providing an additional analytical system capabilities and flexibility.

1.3. Certainty level scales

Philosophically speaking, “certainty, like perfection, does not really come in degree, but we can approach it by degrees” (Klein, 1998) for better understanding. But how can certainty be further sub-divided into degrees? And how many levels are appropriate? While there is an on-going discussion in pragmatic and discourse literature on whether epistemic modality markers should be arranged on a continuum or in discrete categories, there seems to be an agreement that there are at least three articulated points on a presumed continuum from certainty to doubt.

Holmes (1982) offers an abstract scale of certainty or likelihood with three established concrete points: certain (e.g., “Inevitably the Polish Pope has drawn attention to himself”); probable (e.g., “I doubt that she’s coming now”); and possible (e.g., “They might keep them in the backroom”). Hoye (1997) adopts this epistemic trichotomy of certainty, probability, and possibility, also without distinctions for certainty of assertions or negations.

In attitude and affect natural language computing, in the context of extracting opinions from news article corpora, Rubin and colleagues (2004, 2005) extended Hoye–Holmes models by adding two extremes on the epistemic continuum scales:
absolute certainty (defined as a stated unambiguous indisputable conviction or reassurance) and uncertainty (defined as hesitancy or stated lack of clarity or knowledge), and re-defined the middle categories as high certainty (i.e., high probability or firm knowledge), moderate certainty (i.e., estimation of an average likelihood or reasonable chances), and low certainty (i.e., distant possibility, see Fig. 1).

While Rubin (2006) model is primarily concerned with identification of certainty levels encoded in explicit certainty markers in propositions, it also takes into account three contextual dimensions relevant to news discourse. Perspective attributes explicit certainty either to the writer or two types of reported sources – direct participants and experts in a field. Focus separates certainty in facts and opinions. Time is an organizing principle of news production and presentation, and if relevant, is separated into past, present, or future.

Having compared two uncertainty types relevant to IS situations (the cognitive, affective, experiential type and the linguistic content-based epistemic modality), the present study further presents the results of the manual annotation and tests the viability of the discrete categories hypothesized in Rubin (2006) conceptual model for five levels on epistemic continuum and three pragmatic dimensions (perspective, focus, and time) in the context of information seeking from news reports and editorial (Fig. 1).

2. Methodology

The dataset of 80 articles (40 editorials and 40 news reports) was randomly selected from The New York Times articles published in 2000. It consisted of a total of 2243 sentences, with 866 sentences in the editorials and 1377 sentence in the news reports (Rubin, 2006). A subset of 10 articles (272 sentences, about 12% of the full dataset) was analyzed by four independently trained annotators (excluding the author). The agreement results were evaluated in two consecutive intercoder reliability experiments.

2.1. Data analysis

Each sentence was subject to content and linguistic analyses by the author. Each statement in a news article was considered to be a potential locus of explicit certainty. Each explicit certainty marker was assigned a certainty level and placed in its pragmatic contexts (see D2–D4, Fig. 1). Each marker was only assigned one category from each dimension.

Fig. 1. The model for uncertainty–certainty continuum in three news contexts for information seeking as interaction with texts (redrawn from Rubin, 2006).
The number of occurrences of markers per article was totaled and divided by the article sentence length, resulting in one certainty measure – a frequency score per article. The counts of certainty markers per sentence across newspaper articles were summarized in terms of two sample groups, the news reports and the editorials. Within-group and across-group variance were tested with an independent measures two-tailed t-test.

2.2. Annotation process

The manual annotation scheme was defined in the codebook instructions that specified the annotation procedures for determining certainty-qualified statements, the order of assigning categories, and exemplified each certainty category (Rubin, 2006). In Experiment 1, three coders received individual one-hour training regarding the use of the codebook, originally written in a general suggestive tone. In Experiment 2, the fourth annotator went through a more thorough 5-h training and used a revised, more rigidly-specified codebook with an alphabetized key-word index.

2.3. Intercoder agreement measures

Each pair of coders were evaluated on whether they agreed regarding (1) the sentences that contained explicit certainty markers; (2) the specific certainty markers within agreed-upon certainty-qualified sentences; and (3) the classification of agreed-upon markers into one of the categories within each dimension (i.e., level, perspective, focus and time). The sentence and marker agreement measures were calculated with percent agreement. The agreed-upon marker category assignments were assessed in each pair of independent coders with Cohen’s kappa statistic (Cohen, 1960), averaged, and compared to the author’s annotation.

3. Results and discussion

3.1. Certainty distribution in news context

In the overall dataset of 2243 sentences, 1330 sentences (59%) were identified as epistemically modalized, and 913 sentences (41%) were unmarked, which means that there was a ratio of approximately three certainty-qualified statements to two unmarked statements. An average frequency score was 0.82 explicit certainty markers per sentence (with the normalization per article sentence length). The editorial sample group contained significantly more explicit certainty markers per sentence ($M = 0.94, SD = 0.25$) than the news reports sample group ($M = 0.7, SD = 0.29; p = 0.0002$, two-tailed t-test). This means that explicitly certainty-qualified, or epistemically modalized, statements occur at a significantly higher rate in editorials than in news reports, the results are consistent with the pilot data analysis by Rubin and colleagues (2004). In addition, editorials have a high likelihood of starting and ending with an explicit certainty-qualified statement, while news reports tend to start with an implicitly certain statement, and have equal chances of ending with an explicitly certainty-qualified statement.

3.2. Relative category distributions of certainty markers

In the analysis of individual category distributions within each dimension, high and moderate certainty levels were found to be the two most highly represented certainty levels (Fig. 2). High certainty in the news reports is most frequent (55% of
total number of all identified certainty markers); uncertainty in editorials is the least (only 3%). Editorials contain 83% of certainty from the writer’s point of view; while in the news reports the certainty from the writer’s point of view is almost equally distributed with the certainty from the reported direct participants’ point of view. Varying levels of certainty are expressed more in opinions than in factual statements regardless of the article type (news or editorials). The time with which certainty is expressed was distributed in almost equal parts between past, present, and future, with time being rarely irrelevant in its pragmatic context. How these results could be most useful in the IS context is yet to be explored.

3.3. Typology of certainty markers

In 1330 explicitly certainty-qualified sentences there were 1727 occurrences of markers of various epistemic strength. They were grouped into a typology of 43 syntactico-semantic classes; each class is likely to occur within one of the five levels of certainty (Rubin, 2006). Among the most frequently used certainty markers are central modal auxiliary verbs (e.g., must, could), gradable adjectives in their superlative degree, and adverbial intensifiers (e.g., much and so), while adjectival down-toners (e.g., feebly + NP) and adverbial value disjuncts (e.g., annoyingly, rightly) are rarely used to express explicit certainty. The typology will become a basis for an automated certainty identification algorithm with an intention of improving IS systems for intelligence and business analysts that deal with news articles on a daily basis.

3.4. Intercode reliability test results

In Experiment 1, the three coders agreed on whether a statement was modalized by an explicit certainty marker or not 71% of the time with 0.33 Cohen kappa, on average. Within agreed-upon certainty-qualified sentences, they agreed on actual certainty markers, on average, 54% of the time. In the categorization task for the agreed-upon markers, the three coders, on average, were able to reach a slight agreement in the level and focus dimensions (0.15 and 0.13 kappa statistics, respectively), and a fair agreement in perspective and time dimensions (0.44 and 0.41 kappa) according to the Landis and Koch (1977) agreement interpretation scale. The subsequent Experiment 2 showed promising results in agreement on explicit certainty markers (67%) and overall ability to distinguish certainty-qualified statements from unmarked statements (0.51 kappa), and in the relatively intuitive categorization of the perspective dimension (0.65 kappa).

Although stricter instructions may have imposed a more orderly way of looking at the epistemic continuum, the five level certainty boundaries are still subject to individual perceptions (0.41 kappa in Experiment 2). A possibly explanation is that people might be using the same expressions but their underlying categorization systems may not overlap accurately or be interpreted in the same way. This is consistent with Burkell (2004) findings of great variability of interpretations epistemic modality markers, called “verbal labels of likelihood” in her interdisciplinary research on consumer health informatics and cognitive psychology. She emphasizes that while “verbal labels of likelihood are viewed as easy to use, their interpretation is highly variable and dependent on the specific context” (p. 202). Particular meta-cognitive difficulties arose with categorization of negations. For instance, a phrase “I’m sure I don’t know” may be annotated as either absolute certainty or complete uncertainty.

Alternatively, in spite of its large inventory of certainty markers, English may not be precise enough to reliably distinguish multiple epistemic shades. Recent pragmatic, discourse, and philosophy of language studies in mood and modality call for more comprehensive and truer to natural language description of epistemic modality in English reference grammar materials (Hoye, 2005). The latest modality scholarship will undoubtedly contribute to natural language computing applications with important implications for IS systems design.

Time categorization in the context of certainty remained a challenge in spite of more vigorous training in Experiment 2 (0.31 kappa). The interpretation of the reference point of “the present” in the reported speech and nested events can be ambiguous in the certainty identification task. Distinguishing facts versus opinions in combination with certainty identification also presented a particularly puzzling cognitive task (0.16 kappa), possibly due to necessity to evaluate closely related facets of a statement: whether the statement is purely factual, and how sure the author is about the proposition. The possibility of epistemically modalized facts is particularly intriguing.

4. Conclusions

In an effort to provide an interdisciplinary bridge between information seeking, linguistics, natural language processing, and information retrieval, this research (1) introduced a type of uncertainty that is a characteristic of textual information and requires interpretation by information seekers; (2) shifted attention from uncertainty to the epistemic continuum with gradients from uncertainty to absolute certainty; (3) reported empirical results of the manual annotation of texts in American English written news discourse; and (4) tested how much agreement independent users can reach in decoding discrete levels of certainty on the epistemic continuum.

The study found an extensive presence of epistemic modality markers in American English texts, with a significantly higher rate of occurrence of epistemically modalized statements in editorials than in news reports. The most prominent patterns and regularities were identified, and the linguistic means of expressing varying levels of certainty were documented and arranged into the typology of syntactico-semantic classes (further details in Rubin, 2006). This study implies that boundaries between shades of certainty on an epistemic continuum (such as probability and possibility) are highly subjective and
present difficulties in manual annotation. Several explanations for these modest intercoder reliability results were offered: a complexity of the cognitive task of distinguishing five gradients of epistemic modality, a variability in individual user’s underlying categorization systems for levels of estimation of likelihood, or the inherent ambiguity of the English inventory of the epistemic modality means. It is important to note that all four independent annotators were able to recognize the existence of gradients on the continuum, and that stricter annotation instructions and longer coder training improved intercoder agreement results.

Continued research into fine-tuning the number of perceived categories on the epistemic continuum and connecting them to other pragmatic evidence in different discourse contexts can contribute to the flexibility of information systems with identifying levels of epistemic modality or retrieving information with desired epistemic strength (DiMarco, Mercer, & Rubin, 2006). Lastly, I invite information seeking researchers to re-evaluate text interactions in the IS context in this light of linguistic awareness about levels of epistemic modality that are inherently expressed in texts.

Acknowledgements

I would like to thank Dr. Jacqueline Burkell and Dr. Pam McKenzie at the University of Western Ontario, and my doctoral committee members at Syracuse University for their thoughtful suggestions, as well as the four anonymous annotators from Michigan State University for their time and energy put into the annotation process.

References

