Stating with Certainty or Stating with Doubt: Intercoder Reliability Results for Manual Annotation of Epistemically Modalized Statements

Victoria L. Rubin
Faculty of Information and Media Studies
University of Western Ontario
London, Ontario, Canada N6A 5B7
vrubin@uwo.ca

Abstract

Texts exhibit subtle yet identifiable modality about writers' estimation of how true each statement is (e.g., *definitely true* or *somewhat true*). This study is an analysis of such explicit certainty and doubt markers in epistemically modalized statements for a written news discourse. The study systematically accounts for five levels of writer's certainty (ABSOLUTE, HIGH, MODERATE, LOW CERTAINTY and UNCERTAINTY) in three news pragmatic contexts: perspective, focus, and time. The study concludes that independent coders' perceptions of the boundaries between shades of certainty in epistemically modalized statements are highly subjective and present difficulties for manual annotation and consequent automation for opinion extraction and sentiment analysis. While stricter annotation instructions and longer coder training can improve intercoder agreement results, it is not entirely clear that a five-level distinction of certainty is preferable to a simplistic distinction between statements with certainty and statements with doubt.

1 Introduction

1.1 Epistemic Modality, or Certainty

Text conveys more than just a writer's propositional context of assertions (Coates, 1987), e.g., *X 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., *X must be true*, *Y thinks that X is true*, or *perhaps X is true*. A statement is qualified in such a way (beyond its mere referential function) is modal, or epistemically modalized (Coates, 1987; Westney, 1986).

CERTAINTY, or EPISTEMIC MODALITY, concerns a linguistic expression of an estimation of the likelihood that a certain state of affairs is, has been, or will be true (Nuyts, 2001). Pragmatic and discourse 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 (Lackoff, 1972) and hedging in news writing (Hyland, 1999; Zuck & Zuck, 1986). Little attempt, however, has been made in natural language computing literature to 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 stands to benefit the tasks of

Proceedings of NAACL HLT 2007, Companion Volume, pages 141–144,
Rochester, NY, April 2007.

opinion and sentiment analysis and extraction in natural language computing.

1.2 Certainty Level Scales

While there is an on-going discussion in pragmatic 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. Hoye (1997) suggested an epistemic trichotomy of CERTAINTY, PROBABILITY, and POSSIBILITY, consistent with Holmes' (1982) scale of certainty of assertions and negations where the writer asserts WITH CERTAINTY that a proposition is true or not true; or that the proposition is PROBABLY or POSSIBLY true or not true. In attitude and affect computational analysis literature, 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).

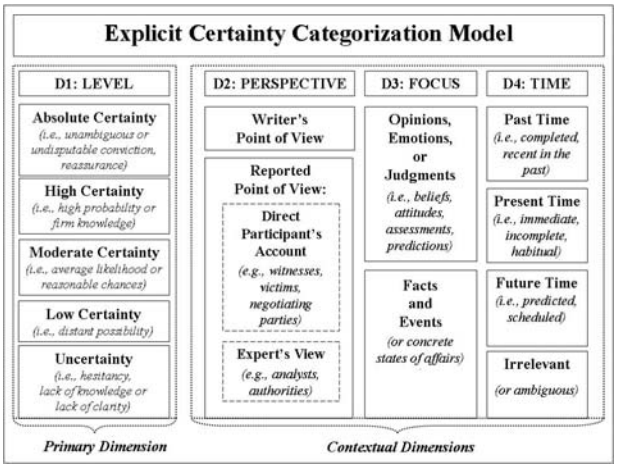

Figure 1. Revised Explicit Certainty Categorization Model (redrawn from Rubin, 2006).

While Rubin's (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.

2 Methodology

This study uses the above-described conceptual certainty categorization model to annotate a news dataset, and produce a typology of syntactic, semantic and lexical classes of certainty markers that map statements into 5 levels of certainty ranging from absolutely certain to uncertain.

The dataset consisted of 80 randomly selected articles (from the AQUAINT Corpus of English Texts, distributed by The New York Times Services in 2000). It constituted a total of 2,243 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 4 independently trained annotators (excluding the author). The agreement results were evaluated in 2 consecutive intercoder reliability experiments.

2.1 Annotation Process

The manual annotation scheme was defined in the codebook instructions that specified the 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 annotation scheme, and were instructed to use the original codebook written in a general suggestive tone. In Experiment 2, the fourth annotator went through a more thorough five-hour training and used a revised, more rigidly-specified codebook with an alphabetized key-word index mapped certainty markers into 5 levels.

Each statement in a news article (be it a sentence or its constituent part such as a clause) was a potential locus of explicit certainty. In both experiments coders were asked to decide if a sentence had an explicit indication of a certainty level. If so, they then looked for explicit certainty markers that contributed to that indication. If a sentence contained a certainty marker, the annotators were in-

structed to consider such a sentence certainty-qualified. The statement was assigned a certainty level and placed in its pragmatic context (i.e., into one of the categories) within the perspective, focus, and time dimensions (see D2 – D4, Fig. 1) relevant to the news discourse. Each marker was only assigned one category from each dimension.

2.2 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) classification of the 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. Partial word string matches were considered a marker match but were weight-adjusted. 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 Typology of Certainty Markers

The content analysis of the dataset generated a group of 1,330 explicitly certainty-qualified sentences with 1,727 occurrences of markers. The markers were grouped into a typology of 43 syntactico-lexical classes; each class is likely to occur within one of the 5 levels of certainty. The typology will become a basis for an automated certainty identification algorithm. 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 downtoners (e.g., *feeble* + NP) and adverbial value disjuncts (e.g., *annoyingly*, *rightly*) are rarely used to express explicit certainty.

3.2 Intercoder Reliability Test Results

In Experiment 1, 1) three coders agreed on whether a sentences was modalized by an explicit certainty marker or not 71% of the time with 0.33 Cohen's kappa, on average. 2) Within agreed-upon certainly-qualified sentences, three coders agreed on actual certainty markers 54% of the time, on average, based on a combined count of the full and weight-adjusted partial matches. 3) 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 5 level certainty boundaries are still subject to individual perceptions (0.41 kappa) and may present difficulties in automation. In spite of its large inventory of certainty markers, English may not be precise enough to reliably distinguish multiple epistemic shades between certainty and doubt. Alternatively, people might be using same expressions but underlying categorization systems for different individuals do not overlap accurately. 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 applications such as opinion extraction and sentiment analysis.

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 and Applications

This study reported the results of the manual annotation of texts in written news discourse, and identified the most prominent patterns and regularities in explicitly stated markers occurrences in modalized statements. The linguistic means of expressing varying levels of certainty are documented and arranged into the typology of syntactico-semantic classes. This study implies that boundaries between shades of certainty in epistemically modalized statements (such as probability and possibility) are highly subjective and present difficulties in manual annotation. This conclusion may warrant a simplification of the existing 5 certainty levels to a basic binary distinction between certainty and doubt. A baseline for future attempts to improve the calibration of levels and their boundaries was established. These modest intercoder reliability results attest to the complexity of the automation of the epistemically modalized statements ranging from certainty to doubt.

In the future studies, I intend to revise the number of the discrete categories on the epistemic continuum and further re-define certainty levels conceptually. I plan to further validate the collection of agreed-upon certainty markers on a much larger dataset and by using the typology as input data to machine learning algorithms for certainty identification and extraction.

References

Coates, J. (1987). Epistemic Modality and Spoken Discourse. *Transactions of the Philological Society*, 110-131.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37-46.

Holmes, J. (1982). Expressing Certainty and Doubt in English. *RELC Journal, 13*(2), 9-29.

Hoye, L. (1997). *Adverbs and Modality in English*. London, New York: Longman.

Hoye, L. (2005). "You may think that; I couldn't possibly comment!" Modality Studies: Contemporary Research and Future Directions. Part II. *Journal of Pragmatics, 37*, 1481-1506.

Hyland, K. (1999). Academic attribution: Citation and the construction of disciplinary knowledge. *Applied Linguistics, 20*(3), 341-367.

Lackoff, G. (1972). *Hedges: a study of meaning criteria and the logic of fuzzy concepts.* Paper presented at the Chicago Linguistic Society Papers.

Landis, J., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159-174.

Light, M., Qiu, X. Y., & Srinivasan, P. (2004). *The Language of Bioscience: Facts, Speculations, and Statements in Between.* Paper presented at the BioLINK 2004: Linking Biological Literature, Ontologies, and Databases.

Mercer, R. E., DiMarco, C., & Kroon, F. W. (2004). *The Frequency of Hedging Cues in Citation Contexts in Scientific Writing.* Paper presented at the Proceedings of the 17th Conference of the CSCSI/SCEIO (AI'2004).

Mushin, I. (2001). *Evidentiality and Epistemological Stance: Narrative Retelling* (Vol. 87). Amsterdam, Philadelphia: John Benjamins Publishing Company.

Nuyts, J. (2001). *Epistemic Modality, Language, and Conceptualization: A cognitive-pragmatic prespective* (Vol. 5). Amsterdam, Philadelphia: John Benjamin Publishing Company.

Palmer, F. R. (1986). *Mood and Modality*. Cambridge: Cambridge University Press.

Rubin, V. L. (2006). *Identifying Certainty in Texts.* Unpublished Doctoral Thesis, Syracuse University, Syracuse, NY.

Rubin, V. L., Kando, N., & Liddy, E. D. (2004). *Certainty Categorization Model.* Paper presented at the AAAI Spring Symposium: Exploring Attitude and Affect in Text: Theories and Applications, Stanford, CA.

Rubin, V. L., Liddy, E. D., & Kando, N. (2005). Certainty Identification in Texts: Categorization Model and Manual Tagging Results. In J. Wiebe (Ed.), *Computing Attitude and Affect in Text: Theory and Applications (The Information Retrieval Series)*: Springer-Verlag New York, Inc.

Westney, P. (1986). How to Be More-or-Less Certain in English - Scalarity in Epistemic Modality. *IRAL: International Review of Applied Linguistics in Language Teaching, 24*(4), 311-320.

Wiebe, J., Wilson, T., & Cardie, C. (2005). *Annotating Expressions of Opinions and Emotions in Language*. Netherlands: Kluwer Academic Publishers.

Zuck, J. G., & Zuck, L. V. (1986). *Hedging in Newswriting. beads or braclets? How do we approach LSP?* Paper presented at the Fifth European Symposium on LSP.