Certainty Categorization Model

Victoria L. Rubin*, Noriko Kando, Elizabeth D. Liddy***

*School of Information Studies
Center for Natural Language Processing
Syracuse University
Syracuse, NY13244-1190, USA
{vlrubin, liddy}@syr.edu

**National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku
Tokyo 101-8430, Japan
kando@nii.ac.jp

Abstract

We present a theoretical framework and preliminary results for manual categorization of explicit certainty information in 32 English newspaper articles. The explicit certainty markers were identified and categorized according to the four hypothesized dimensions – perspective, focus, timeline, and level of certainty. One hundred twenty one sentences from sample news stories contained a significantly lower frequency of markers per sentence (M=0.46, SD =0.04) than 564 sentences from sample editorials (M=0.6, SD =0.23), p= 0.0056, two-tailed heteroscedastic t-test. Within each dimension, editorials had most numerous markers per sentence in high level of certainty, writer's point of view, and future and present timeline (0.33, 0.43, 0.24, and 0.22, respectively); news stories – in high and moderate levels, directly involved third party's point of view, and past timeline (0.19, 0.20, 0.24, and 0.20, respectively). These patterns have practical implications for automation. Further analysis of editorials showed that out of 72 combinations possible under the hypothesized model, the high level of certainty from writer's perspective expressed abstractly in the present and future time, and expressed factually in the future were very common. Twenty two combinations never occurred; and 35 had ≤ 8 occurrences. This narrows the focus for future linguistic analysis of explicit certainty markers.

Introduction

Certainty identification presents an ongoing challenge in information extraction. The notion of certainty, or uncertainty, falls under the speculative type of subjectivity (Wiebe 2000). Subjectivity has been defined as "aspects of language used to express opinions and evaluations" (Banfield 1982, cited in Wiebe 1994, 2000, Wiebe et al. 2001). Subjectivity tagging is considered particularly relevant for the news report genre (Wiebe et al. 2001).

Liddy et al. (1993) applied news report schemata components for an automated text structurer, and noted that subjectivity, or objectivity, as an attribute in texts deserved special attention. Two important observations were made: first, binary distinctions (e.g. +subjective, -subjective) may not be sufficient to adequately represent micro-level similarities and distinctions in texts; and second, discourse components may have multiple dimensions embedded in each of the concept labels (Liddy et al. 1995). This study explores dimensions of certainty in written texts.

There are *three areas of confusion* related to the concept of certainty as a subjective evaluation of texts. *First,* by analogy with subjectivity, certainty is not a grammatical feature but rather a pragmatic position. Subjectivity, a metaphor for "a point of view" or "an angle of vision", originally borrowed from the visual arts, is a spatial notion by nature, and in language, it is taken to be located in a speaker (Banfield, 1982). Expressing certainty or uncertainty in written texts is inevitable, just as one is bound to have a spatial angle of vision. Thus, each statement should potentially reveal a particular certainty.

The second source of confusion is related to distinguishing the writer's certainty as expressed in text, and the reader's certainty that the text is believable. The writer's certainty about his or her own and others' assertions is captured in texts. The reader's certainty is related to the numerous factors which inform his or her own subjectivity, or point of view. The former is accessible for analysis since it has a written record, but the latter is less tangible and may reflect high inter-personal variability. Thus, the reader's certainty is out of scope for this study. The study focuses on the writer's certainty model and its multi-dimensional complexity in the newspaper context.

The third source of confusion affecting certainty tagging is a lack of precision in certainty definitions, which usually revolve around the notions of "*the quality or state of mind of being free from doubt, especially on the basis of evidence"* (Merriam-Webster 2004). In the context of news communications, two basic relational categories are undefined: *whose mind* is free from doubt about *what*.

Some writers consciously strive to produce a particular effect of certainty due to training or overt instructions. Others may do it inadvertently. Writer's certainty level may remain constant in a text and be unnoticed by the reader, or it may fluctuate from statement to statement and blatantly attract readers' attention. There may be evident traces of such writers' behavior that may become apparent

upon a closer examination with a systematic theoretical framework. The difficulty is to discern such traces at the discourse, syntactic, and semantic levels, wherever such explicit information is available and to be able to recognize these explicit markers with a series of NLP algorithms.

The necessity to clarify the areas of confusion and to more clearly define the notion of certainty gave rise to the development of a theoretical categorization model that depicts certainty along four dimensions. In the remainder of the paper, we discuss the model, report on preliminary results, and conclude with outlined challenges and applications.

Proposed Certainty Model

Working Definition of Certainty. For information extraction purposes, we extended the initial dictionary definition of certainty to include its relational characteristics:

> *Certainty* is the quality or state of being free from doubt, especially on the basis of evidence about the past, present, or future factual or abstract information, expressed by the writer or reported by the writer about others, directly or indirectly involved in the events in the narrative.

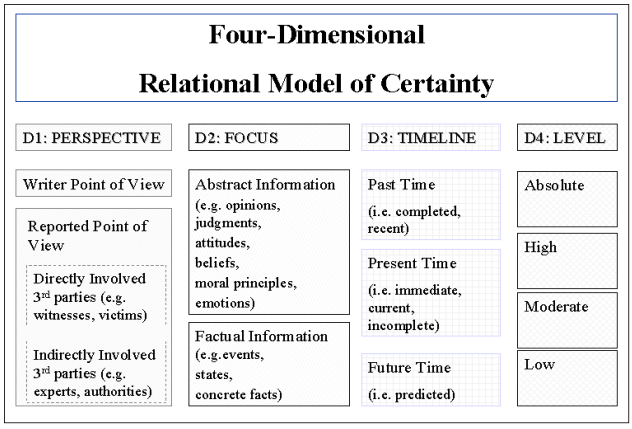

Figure 1. Graphical representation of the concepts of four hypothesized dimensions (across) and their categories (down).

First Dimension: Certainty Perspective. The first dimension in Figure 1, the certainty perspective, separates the certainty point of view into the writer's and the reported points of view. A practical question is whether third parties' voices can be isolated from the author's since they are presented through the author's prism. Reported point of view comprises two groups. First, those of directly involved third parties, such as victims, witnesses, and survivors, are direct event participants, who are either present at the described event or whose life is directly effected by the events. Second, those of indirectly involved third parties, such as experts, authorities, and analysts, who are tangentially related to the event in professional or other capacities.

Second Dimension: Certainty Focus. The certainty focus is divided into abstract and factual information in the narrative. We use *focus* in van Dijk's (1981) localized selection sense as a referent, *viz.* the object, subject, or topic of conversation that is being talked about, or predicated upon in a particular localized syntactic unit, such as a sentence or clause. Abstract information may include judgments, opinions, attitudes, beliefs, moral principles, and emotions. Factual information contains reports of state or events, evidence, and known facts.

Third Dimension: Timeline. The third dimension accounts for relevance of time (past, present and future) to the moment when the article was written. The past naturally includes completed or recent states or events; the present is current, immediate, and incomplete states of affairs; and the future is predictions, plans, warnings, and suggested actions.

Fourth Dimension: Certainty Levels. The concept of certainty seems to fall inherently into levels. Our current model suggests the distinction into four categories - absolute, high, moderate, and low. We assume that the commonly used declarative mood of stating facts and opinions does not contain explicit indication of certainty.

Research Questions

The study will empirically determine:

1) if the sample data support the hypothesized four dimensional categorization model,
2) if so, which categories are most and least frequent for a sample of English news articles,
3) if the data do not support the model, how the categorization might be enhanced;
4) whether there are differences in certainty distributions between editorials and news stories, overall and per hypothesized category;
5) how many perceived categories of certainty can be distinguished within each dimension.

Data and Analysis Methods.

We manually analyzed 32 articles published in the first week of January 2000 (from the AQUAINT Corpus of English Texts), a total of 685 sentences, excluding headlines. The topics of the sample articles varied - the editorials included discussions of political leaders, presidential and state government campaigns, the economic and financial situations in US, Croatia, and Angola, recent historical discoveries, pharmaceutical consumer alerts, and role of the Internet and computers in everyday lives. The news included reports on the misnumbering of New York Times issues, on the

controversy around the millennium and Y2K bug, and women's basketball.

The data were analyzed manually at the sentence-level by one coder, the first author. If a sentence contained explicit certainty information markers, it was decomposed along each certainty dimension by answering questions such as "What is the certainty level?" and "Whose perspective is being presented?" The number of occurrences of markers per article were totaled and adjusted for article sentence length, resulting in one frequency score per article. The length of explicit certainty markers was not pre-determined.

First, we were interested in an overall frequency of occurrence of explicit certainty markers across all of the data. Second, we identified whether the two sample groups had significantly different means. Third, we looked at the overall distribution of frequencies scores (in markers per sentence) per category within each dimension. For instance, were there more occurrences of high or low levels of certainty on average? Fourth, for the editorial sample, we identified the least and most frequent combinations out of 72 possible dimension-category combinations. And last, we assessed whether the data easily fell into the hypothesized categories.

Results and Discussion.

In the total set of 32 articles (685 sentences), an average of 0.53 explicit certainty markers per sentence were identified. Identified certainty markers included but were not limited to *it was not even clear that*, *remains to be seen*, *don't believe they will*, *not necessarily*, *we thought*, *estimated*, *seems exaggerated, would probably have to, is expected to,* and *will almost certainly have to.*

The sample group of 28 editorials (564 sentences, *M*=20, *SD*=5.34 sentences) contained more explicit certainty markers per sentence (*M*=0.6, *SD* =0.26) than the sample group of 4 news stories (121 sentences, *M*=26, *SD*= 8.01 sentences; *M*=0.46, *SD* =0.04 markers per sentence). This difference was statistically significant, p = 0.0056, two-tailed heteroscedastic t-test.

Within each dimension, average frequencies of occurrence of explicit certainty markers per sentence differed from category to category in the level, perspective, and timeline dimensions, as well as between sample groups. Table 1 shows that, overall, out of all possible levels, the high certainty level contained most markers per sentence (0.33). Here is an example from this group[1]:

The crowd cheering the opening of the Erie Canal in 1824 ***knew*** *that the city* ***would forever be***[2] *transformed, Wallace notes. (ID=e28.19::* ***<high level>*** *<directly involved third party's > <factual information> < in the past>)*

In news stories, both high and moderate levels of certainty were the two most prominent levels (approximately 0.2 markers per sentence). An example of the moderate level of certainty follows:

But as midnight closed in, the streets teemed with people ***and there seemed to be little left of the anxiety over terrorist attacks*** *that prompted the mayor of Seattle last week to cancel a major outdoor celebration around the city's famed Space Needle. (ID=n3.9:: <***moderate level***> <writer's> <factual assessment of emotional state in the past>)*

Dimension	Certainty Level			
Sample Group Statistic	**Absol**	**High**	**Mod**	**Low**
editorials *M*, markers per sent.	0.07	**0.33**	0.17	0.04
SD	0.09	0.17	0.14	0.06
news stories *M*, markers per sent.	0.03	**0.19**	**0.20**	0.04
SD	0.05	0.09	0.15	0.05

Table 1. Comparative distributions of two sample groups' means and standard deviations of markers per sentence in 4 categories within certainty level dimension.

Table 2 demonstrates that in editorials certainty from the writers' points of view is expressed more commonly than certainty of third parties, as is expected. Consider that even though this example sentence talks about a third party, the expressed certainty actually belongs to the writer:

He also ***ought to urge*** *France and Russia to persuade Saddam Hussein to accept the resolution. (ID=e8.14:: <absolute>* ***<writer's>*** *<abstract emotional call for action> <in the future>)*

Dimension	Perspective		
Sample Group Statistic	**Writer's Point of View**	**3rd Directly Involved Party's Point of View**	**3rd Indirectly Involved Party's Point of View**
editorials *M*, markers per sent.	**0.43**	0.13	0.04
SD	0.23	0.13	0.06
news stories *M*, markers per sent.	0.16	**0.24**	0.05
SD	0.10	0.11	0.06

Table 2. Comparative distributions of two sample groups' means and standard deviations of markers per sentence in 3 categories within certainty perspective dimension.

[1] Each example is followed by its unique identification number from the raw data file, and the identified level, perspective, focus, and timeline of certainty associated with the example.

[2] Each example contains a certainty marker highlighted in bold.

We also observed that in news stories attention shifts to the certainty of the directly involved third parties such as presidential candidates, political leaders, a Cuban orphan and his family, and just a person waiting for a flight at the airport whose direct words are cited below:

*"**I think it will probably be OK**…" (ID=n4.23:: <low level> <directly involved third party> <abstract uncertainty> <about future>)*

The indirectly involved third parties are rather rare and usually occur in the form of experts' opinions, sometimes cited as well. For instance, economists' points of view rendered below reflect their certainty, and the writer may or may not be sure about that statement:

***Most economists believe** Alan Greenspan is more responsible for the economy's spectacular performance than Congress, Presidents Bush and Clinton or any other identifiable factor. (ID=e9.1:: <high level> <indirectly involved third party's> <abstract assessment> <in present>)*

Sometimes the reference to the source is vague but it is quite clear that the expressed certainty is writer's:

*Although **some research suggests that** some supplements **can produce** positive health effects, there have also been cases where people have been made ill by supplements, or their conditions have become worse…(ID=e28.3:: <moderate level> <**writer's**> <abstract conviction> <in present>)*

Table 3 reveals that abstract and factual foci of certainty were approximately evenly distributed in both sample groups.

Dimension	**Focus**		**Timeline**		
Sample Group Statistic	**Ab-stract**	**Fact-ual**	**Past**	**Pre-sent**	**Fu-ture**
editorials *M*, markers per sent.	0.33	0.27	0.14	**0.24**	**0.22**
SD	0.20	0.19	0.12	0.18	0.16
news stories *M*, markers per sent.	0.23	0.23	**0.20**	0.11	0.14
SD	0.05	0.09	0.11	0.05	0.09

Table 3. Comparative distributions of two sample groups' means and standard deviations of markers per sentence in focus and timeline dimension categories.

As for the timeline, it is not surprising that news stories have the tendency to report events in the past was, as was captured in the certainty information as well. Editorials' tendency to state opinions about current and predicted events also became apparent. Compare examples from the two different samples:

*The failure lasted only about 30 minutes and had no operational effect, the **FAA said**, adding **that it was not even clear that the problem was caused** by the date change. (ID=n4.19:: <low level> <third indirectly involved party> <factual> <**in the past**>)*

*Whatever happens next, these candidates have shown that one-on-one debates **really can give voters a choice** on issues and on leadership temperament as well. (ID=e16.18:: <high level> <writer's> <abstract assessment> <**in the present**>)*

Many editorials had a closing statement usually in the last sentence containing some certainty markers, for instance, expressing predictions or suggesting actions as below:

***There will be problems** along the way, but **the Internet will likely change the way** America does business far beyond the habits of holiday shoppers. (ID=e2.22:: <high level> <writer's> <**abstract prediction**> <into the future>)*

Table 4 shows the distribution of occurrences of explicit certainty markers for combinations of the categories from the four dimensions in the editorial sample. For instance, absolute level of writer's certainty about abstract information in the past only happened once, while in the present it occurred 18 times. The table forms 72 possible combinations (3 perspectives by 2 foci by 3 timelines by 4 levels), and an additional category that was recorded as containing "none" of the explicit certainty information.

Twenty two combinations had no representation in our data. For instance, directly involved third parties' low level of certainty about abstract information in either past, present or future were never found. Thirty five combinations were found to be rare, with ≤ 8 occurrences

		Level								
		Abst				**Fact**				Grand Total
Per-spect	**Time**	Abs	High	Mod	Low	Abs	High	Mod	Low	
Wr	p	1	8	10	1		12	11		43
	pr	18	**29**	16	8		13	10	1	95
	f	13	**25**	12	2	2	**27**	12	3	96
3rd Dir	p		8	4		1	11	2		26
	pr	2	3	2		1	7	5	1	21
	f	1	8	1		1	11	2	2	26
3rd In-dir	p		3	1	2		4			10
	pr		2	4			4	2	1	13
	f						5			5
None										289
Total		35	86	50	13	5	94	44	8	624

Table 4. Count of occurrences within 72 combinations of categories and count of occurrences of sentences with no explicit certainty markers for the sample of 28 editorials.

in our data, for instance, low level of writer's certainty about present or future factual information had 1 and 3 occurrences respectively. Another 12 combinations

accounted for the majority of occurrences and varied between 10 and 18 occurrences.

The remaining 3 categories had an unusually high representation in editorials. The combinations are writer's high level of certainty about abstract information in present or future, such as predictions and current assessments, which had 29 and 25 occurrences respectively. There were also 27 occurrences of writer's future high level certainty factual predictions, namely, stating with high certainty what will happen in the future.

The observed distribution is consistent with the goal of editorials to state opinions, inevitably with different levels of certainty. It can direct us to the combinations that cover the majority of explicit certainty markers for further linguistic analysis and provide guidance in automating the categorization.

The presence of data in each category suggests that the categorization model is viable when applied manually. Now a gold standard and a set of rules can be created for an inter-coder agreement study and further automation of the process. High frequency of explicit certainty markers in some categories emphasizes where linguistic analysis should be concentrated to cover the majority of certainty expression cases.

Another criteria for deciding whether the sample data support the hypothesized model is "ease-of-fit", in other words, whether the data landed naturally or had to be forced into the allotted categories within each dimension. The easiest dimension for categorization was the timeline. The only adjustment that had to be made was an expansion of the notion of present time to include regular or habitual actions. Certainty level categorization could include an additional fifth category of uncertainty. Currently, no distinction between low certainty and uncertainty has been made. The perspective, on the other hand, is sufficiently granular and, depending on application, could even be collapsed into two main categories: the writer's and 3rd party's points of view. The benefit of distinguishing a rather rare category of 3rd indirectly involved party's perspective is for when we are particularly interested in, let's say, experts' certainty. The distinction of focus into factual and abstract (or non-factual) information presented most difficulties for annotation due to fuzzy boundaries between known facts and opinions. The focus was considered factual when an event or state of affairs was clearly mentioned. Otherwise, the focus was considered abstract and further sub-categorized into a type of opinion, judgment, or emotion, such as fear, warning, an assessment, a prediction, or conviction. The annotation could be improved with an explicit set of guidelines and definitions. All of the hypothesized categories in the model are not final and are open to further refinement as the data analysis proceeds and the theoretical framework stabilizes. The first author plans to incorporate some of the above-mentioned refinements into her doctoral thesis.

Challenges

The proposed model makes several assumptions and raises several philosophical and practical issues. For instance, we are assuming that uncertainty is expressed due to doubt on the basis of evidence (by our definition), thus we do not make a distinction between truly being uncertain and appearing to be uncertain. There may be other desired reasons for appearing to be uncertain, such as the psychological effects of non-aggression, social politeness effect, humbling effect of hedged speech, and practical concerns for avoiding liabilities. Identifying these pragmatic functions of uncertainty poses a challenge for future automated identification, and is currently out of scope of the study. Another problem is literal interpretation of the identified clues. For instance, the word "certain" itself has an alternate meaning of "definite but not specified". Our model does not include this meaning, but the issue of contextual disambiguation still persists.

Applications

The categorization, and the resulting linguistic clues and patterns for most frequent categories, will serve as a starting point for a certainty identification module in an intelligence analyst's question and answering system. This model will be applied to identifying and extracting perceived certainty of specified writers or reported third parties relative to topics of interest. For instance, how certain are President Bush's statements when predicting the outcome of the Middle East conflict?

The collection of certainty expressions may become input data to machine learning algorithms for certainty identification and extraction. It also may suggest a new way of automating genre identification.

In addition, the study results capture current trends in newspaper writing, and are potentially useful as a set of suggestions on how to convey a desired level of certainty.

Conclusions and Future Work

Our contribution is in a proposed relational model and analytical framework for certainty categorization. Preliminary results reveal an overall promising picture of the presence of certainty information in texts, and establish the ability to manually identify and categorize individual statements.

Editorials had a significantly higher frequency of markers per sentence than did the news stories. For editorials, high level of certainty, writer's point of view, and future and present timelines were the most populated categories, while for news stories, the most common were high and moderate levels, directly involved third party's point of view, and past timeline. We are interested in conducting further data analysis per genre within

newspaper articles since we have established that the frequency distribution differs depending on genre. This may have implications for automated genre identification. We will use insights from previous work on genre classification (Liddy et al. 1995, Kando 1996).

For editorials, out of possible combinations, the high level of certainty from the writer's point of view expressed abstractly in the present and future time and expressed factually in the future were very common; 35 were rather rare; and 22 combinations never occurred. These results shed light on where the majority of lexical, semantic and syntactic patterns can be expected during linguistic analysis of editorials.

The sample data fit relatively well into the pre-defined categories. Some categories, such as the certainty level, can still be further refined with finer distinctions. The focus dimension will require further research. The study yielded a collection of explicit certainty markers which are to be further grouped and analyzed in terms of lexical, semantic and syntactic patterns.

We also plan to conduct an inter-coder reliability study with multiple annotators by adopting our online data collection facility, developed for a concurrent study of emotional subjective content (Rubin, Stanton, and Liddy. forthcoming).

Acknowledgements

This research was made possible by the National Science Foundation East Asia Summer Institutes for U.S. Graduate Students Research Grant No. 0309745. The first author extends her gratitude to her host researchers, Dr. Kando and Dr. Jun Adachi, for welcoming this effort at the National Institute of Informatics, Tokyo, Japan. We are also grateful to the colleagues at Dr. Nakagawa's Language Informatics Laboratory, Information Technology Center at the University of Tokyo, and the researchers at Dr. Isahara's Computational Linguistics Group at the Communications Research Laboratory in Kyoto, Japan, for their comments and suggestions during the early stages of the research. We would especially like to thank Dr. Janyce Wiebe for her input in personal interactions at the ACL 2003, 41st Annual Meeting of the Association for Computational Linguistics in Sapporo, Japan.

References

Banfield, A. 1982. *Unspeakable Sentences.* Routledge and Kegan Paul, Boston

Kando, N. 1996. Text structure analysis based on human recognition: Cases of Japanese newspaper and English newspaper. *Bulletin of National Center for Science Information Systems*, No. 8, pp.107-126 (Japanese)

Liddy, E. D., McVearry, K., Paik, W., Yu, E.S., McKenna, M. 1993. Development, implementation & Testing of a Discourse Model for Newspaper Texts. *Proceedings of the ARPA Workshop on Human Language Technology,* Princeton, NJ, March 21-24, 1993.

Liddy, E.D., Paik, W., McKenna, M. 1995. Development and Implementation of a discourse model for newspaper texts. *In Proc. of the AAAI Symposium on Empirical Methods in Discourse Interpretation and Generation.* Stanford, CA.

Merriam-Webster Online Dictionary, http://www.m-w.com/ Accessed, January 30, 2004.

Rubin, V.L., Stanton, J.M., Liddy E.D. Discerning Emotions in Texts. Forthcoming..

van Dijk, T. A. 1981. *Studies in the Pragmatics of Discourse*, Mouton Publishers, The Hague, The Netherlands

van Dijk, T. A. 1988. *News Analysis: Case Studies of International and National News in the Press.* Lawrence Erlbaum, Hillsdale, New Jersey

Wiebe, J. M. 1994. Tracking Point of View in Narrative. *Computational Linguistics* 20 (2): 233-287.

Wiebe, J. 2000. Learning Subjective Adjectives from Corpora. *Proc. 17th National Conference on Artificial Intelligence (AAAI-2000).* Austin, Texas, July 2000.

Wiebe, J., Bruce, R., Bell, M., Martin, M., Wilson, T. 2001. A Corpus Study of Evaluative and Speculative Language. *Proc. 2nd ACL SIGdial Workshop on Discourse and Dialogue.* Aalborg, Denmark, September, 2001.

Wilson, T., Wiebe, J. 2003. Annotating Opinions in the World Press. *4th SIGdial Workshop on Discourse and Dialogue (SIGdial-03).* Sapporo, Japan, July 2003.