Levels of Certainty in Knowledge-Intensive Corpora: An Initial Annotation Study

Aron Henriksson
DSV/KTH-Stockholm University
Sweden
aronhen@dsv.su.se

Sumithra Velupillai
DSV/KTH-Stockholm University
Sweden
sumithra@dsv.su.se

Abstract

In this initial annotation study, we suggest an appropriate approach for determining the level of certainty in text, including classification into multiple levels of certainty, types of statement and indicators of amplified certainty. A primary evaluation, based on pairwise inter-annotator agreement (IAA) using F_1-score, is performed on a small corpus comprising documents from the *World Bank*. While IAA results are low, the analysis will allow further refinement of the created guidelines.

1 Introduction

Despite ongoing efforts to codify knowledge, it is often communicated in an informal manner. In our choice of words and expressions, we implicitly or explicitly judge the certainty of the knowledge we wish to convey. This fact makes it possible to gauge the reliability of knowledge based on the subjective perspective of the author.

As knowledge is often difficult to ascertain, it seems reasonable to regard knowledge on a continuum of varying degrees of certainty, as opposed to a binary (mis)conception. This corresponds to the notion of *epistemic modality*: the degree of confidence in, or commitment to, the truth of propositions (Hyland, 1998). *Hedging* is a means of affecting *epistemic modality* by qualifying propositions, realized through tentative words and expressions such as *possibly* and *tends to*.

A holistic perspective on certainty—in which not only speculation is considered, but also signs of increased certainty—requires a classification into various levels. Applying such an approach to knowledge-intensive corpora, it may in due course be possible to evaluate unstructured, informal knowledge. This would not least be valuable to organizational knowledge management practices, where it could provide a rough indicator of reliability in internal *knowledge audits*.

2 Related Research

The *hedging* concept was first introduced by Lakoff (1973) but has only really come into the spotlight in more recent years. Studies have mainly taken place in the biomedical domain, with Hyland's (1998) influential work investigating the phenomenon in scientific research articles. Speculative keywords and negations, along with their linguistic scopes, are annotated in the *BioScope* corpus by Vincze et al. (2008), which contains a large collection of medical and biological text (scientific articles and abstracts, as well as radiology reports). After several iterations of refining their guidelines, they report IAA values ranging from 77.6 to 92.37 F_1-score for speculative keywords (62.5 and 95.54 F_1-score for full scope). This corpus is freely available and has been used for training and evaluation of automatic classifiers, see e.g. Morante and Daelemans (2009). One of the main findings is that hedge cues are highly domain-dependent. Automatic identification of other private states, including opinions, represents a similar task, see e.g. Wiebe et al. (2005). Diab et al. (2009) study annotation of committed and non-committed belief and show that automatic tagging of such classes is feasible. A different annotation approach is proposed by Rubin et al. (2006), in which certainty in newspaper articles is categorized along four dimensions: *level*, *perspective*, *focus* and *time*. Similarly, five dimensions are used in Wilbur et al. (2006) for the creation of an annotated corpus of biomedical text: *focus*, *polarity*, *certainty*, *evidence* and *directionality*.

3 Method

Based on previous approaches and an extensive literature review, we propose a set of guidelines that

(1) incorporates some new features and (2) shifts the perspective to suit knowledge-intensive corpora, e.g. comprising organizational knowledge documents. Besides categorization into levels of certainty, this approach distinguishes between two types of statement and underscores the need to take into account words and expressions that add certainty to a proposition.

A small corpus of 10 *World Bank* documents—a publicly available resource known as *Viewpoints* (The World Bank Group, 2010)—is subsequently annotated in two sets by different annotators. The corpus is from a slightly different domain to those previously targeted and represents an adequate alternative to knowledge documents internal to an organization by fulfilling the criterion of knowledge intensity. The process is carried out in a *Protégé* plugin: *Knowtator* (Ogren, 2006). Pairwise IAA, measured as F_1-score, is calculated to evaluate the feasibility of the approach.

Statements are annotated at the clause level, as sentences often contain subparts subject to different levels of certainty. These are not predefined and the span of classes is determined by the annotator. Furthermore, a distinction is made between different types of statement: statements that give an *account* of something, typically a report of past events, and statements that express concrete knowledge *claims*. The rationale behind this distinction is that text comprises statements that make more or less claims of constituting knowledge. Thus, knowledge *claims*—often less prevalent than *accounts*—should be given more weight in the overall assessment, as the application lies in automatically evaluating the reliability of informal knowledge. Assuming the view of knowledge and certainty as continuous, it is necessary to discretize that into a number of intervals, albeit more than two. Hence, *accounts* and *claims* are categorized according to four levels of certainty: *very certain*, *quite certain*, *quite uncertain* and *very uncertain*. In addition to the statement classes, four indicators make up the total of twelve. We introduce *certainty amplifiers*, which have received little attention in previous work. These are linguistic features that add certainty to a statement, e.g. words like *definitely* and expressions like *without a shadow of a doubt*. *Hedging indicators*, on the other hand, have gained much attention recently and signify uncertainty. The *source hedge* class is applicable to instances where the source of *epistemic judgement* is stated explicitly, yet only when it provides a hedging function (e.g. *some say*). *Modality strengtheners* are features that strengthen the effect of *epistemic modality* when used in conjunction with other (un)certainty indicators—but alone do not signify any polarity orientation—and may be in the form of vagueness (e.g. <could be> *around* that number) or quantity gradations (e.g. *very* <sure>).

4 Results

The corpus contains a total of 772 sentences, which are annotated twice: set #1 by one annotator and set #2 by five annotators, annotating two documents each. The statistics in Table 1 show a discrepancy over the two sets in the number of classified statements, which is likely due to difficulties in determining the scope of clauses. There are likewise significant differences in the proportion between *accounts* and *claims*, as had been anticipated.

Accounts		**Claims**	
Set #1	Set #2	Set #1	Set #2
726	574	395	393

Table 1: Frequencies of accounts and claims.

Despite the problem of discriminating between *accounts* and *claims*, they seem to be susceptible to varying levels of certainty. The average distribution of certainty for *account* statements is depicted in Figure 1. As expected, an overwhelming majority (87%) of such statements are *quite certain*, merely relating past events and established facts.

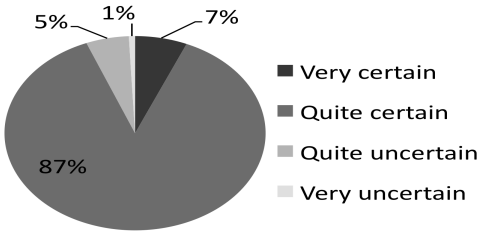

Figure 1: Average distribution of certainty in *account* statements.

By comparison, knowledge *claims* are more commonly hedged (23%), although the majority is still *quite certain*. Interestingly, *claims* are also expressed with added confidence more often than *accounts*—around one in every ten *claims*.

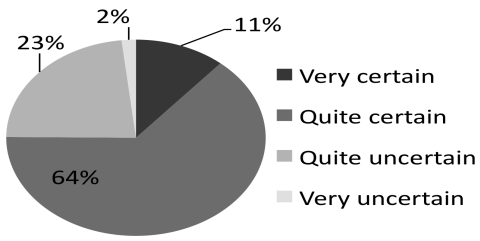

Figure 2: Average distribution of certainty in knowledge *claims*.

As expected, the most common indicator is of hedging. Common cues include *may*, *can*, *might*, *could*, *indicate(s)*, *generally* and *typically*. Many of these cues are also among the most common in the biomedical sub-corpus of *BioScope* (Vincze et al., 2008). It is interesting to note the fairly common phenomenon of *certainty amplifiers*. These are especially interesting, as they have not been studied much before, although Wiebe et al. (2005) incorporate *intensity ratings* in their annotation scheme. There is agreement on words like *clearly*, *strongly* and *especially*.

Indicator	**Set #1**	**Set #2**
Certainty amplifier	61	29
Hedging indicator	151	133
Source hedge	0	40
Modality strengthener	9	122

Table 2: Frequency of indicators

To evaluate the approach, we calculate IAA by pairwise F_1-score, considering set #1 as the gold standard, i.e. as correctly classified, in relation to which the other subsets are evaluated. We do this for exact matches and partial matches[1]. For exact matches in a single document, the F_1-score values range from an extremely low 0.09 to a somewhat higher—although still poor—0.52, yielding an overall average of 0.28. These results clearly reflect the difficulty of the task, although one has to keep in mind the impact of the discrepancy in the number of annotations. This is partly reflected in the higher overall average for partial matches: 0.41.

Certainty amplifiers and *hedging indicators* have F_1-scores that range up to 0.53 and 0.55 respectively (ditto for partial matches) in a single document. Over the entire corpus, however, the averages come down to 0.27 for *certainty amplifiers* (0.30 for partial matches) and 0.33 for *hedging indicators* (0.35 for partial matches).

Given the poor results, we want to find out whether the main difficulty is presented by having to judge certainty according to four levels of certainty, or whether it lies in having to distinguish between types of statement. We therefore generalize the eight statement-related classes into a single division between *accounts* and *claims*. Naturally, the agreement is higher than for any single class, with 0.44 for the former and 0.41 for the latter. A substantial increase is seen in partial matches, with 0.70 for *accounts* and 0.55 for *claims*. The results are, however, sufficiently low to conclude that there were real difficulties in distinguishing between the two.

Statement Type	**Exact F_1**	**Partial F_1**
Account	0.44	0.70
Claim	0.41	0.55

Table 3: Pairwise IAA per statement type, F_1-scores for exact and partial matches.

We subsequently generalize the eight classes into four, according to their level of certainty alone. The results are again low: *quite certain* yields the highest agreement at 0.47 (0.76 for partial matches), followed by *quite uncertain* at 0.24 (0.35 for partial matches). These numbers suggest that this part of the task is likewise difficult. The rise in F_1-scores for partial matches is noteworthy, as it highlights the problem of different interpretations of clause spans.

Certainty Level	**Exact F_1**	**Partial F_1**
Very certain	0.15	0.15
Quite certain	0.47	0.76
Quite uncertain	0.24	0.35
Very uncertain	0.08	0.08

Table 4: Pairwise IAA per certainty level, F_1-scores for exact and partial matches

5 Discussion

In the guidelines, it is suggested that the level of certainty can typically be gauged by identifying the number of indicators. There is, however, a serious drawback to this approach. Hedging indicators, in particular, are inherently uncertain to different degrees. Consider the words *possibly* and

[1]Partial matches are calculated on a character level while exact matches are calculated on a token level.

probably. According to the guidelines, a single occurrence of either of these hedging indicators would normally render a statement *quite uncertain*. Giving freer hands to the annotator might be a way to evade this problem; however, it is not likely to lead to any more consistent annotations. Kilicoglu and Bergler (2008) address this by assigning weights to hedging cues.

A constantly recurring bone of contention is presented by the relationship between certainty and precision. One of the hardest judgements to make is whether imprecision, or vagueness, is a sign of uncertainty. Consider the following example from the corpus:

> *Cape Verde had virtually no private sector.*

Clearly, this statement would be more certain if it had said: *Cape Verde had no private sector.* However, *virtually no* could be substituted with, say, *a very small*, in which case the statement would surely not be deemed uncertain. Perhaps precision is a dimension of knowledge that should be analyzed in conjunction with certainty, but be annotated separately.

6 Conclusion

There are, of course, a number of ways one can go about annotating the level of certainty from a knowledge perspective. Some modifications to the approach described here are essential—which the low IAA values are testament to—while others may be worth exploring. Below is a selection of five key changes to the approach that may lead to improved results:

1. *Explicate statement types*. Although there seems to be a useful difference between the two types, the distinction needs to be further explicated in the guidelines.

2. *Focus on indicators*. It is clear that indicators cannot be judged in an identical fashion only because they have been identified as signifying either certainty or uncertainty. It is not simply the number of occurrences of indicators that determines the level of certainty but rather how *strong* those indicators are. A possible solution is to classify indicators according to the level of certainty they affect.

3. *Discard rare classes*. Very rare phenomena that do not have a significant impact on the overall assessment can be sacrificed without affecting the results negatively, which may also make the task a little less difficult.

4. *Clarify guidelines*. A more general remedy is to clarify further the guidelines, including instructions on how to determine the scope of clauses; alternatively, predefine them.

5. *Instruct annotators*. Exposing annotators to the task would surely result in increased agreement, in particular if they agree beforehand on the distinctions described in the guidelines. At the same time, you do not want to steer the process too much. Perhaps the task is inherently difficult to define in detail. Studies on how to exploit subjective annotations might be interesting to explore, see e.g. Reidsma and op den Akker (2008).

In the attempt to gauge the reliability of knowledge, incorporating multiple levels of certainty becomes necessary, as does indicators of increased certainty. Given the similar rates of agreement on *hedging indicators* and *certainty amplifiers* (0.33 and 0.27 respectively; 0.30 and 0.35 for partial matches), the latter class seem to be confirmed. It is an existing and important phenomenon, although—like *hedging indicators*—difficult to judge. Moreover, a differentiation between types of statement is important due to their—to different degrees—varying claims of constituting knowledge. An automatic classifier built on such an approach could be employed with significant benefit to organizations actively managing their collective knowledge. The advantage of being aware of the reliability of knowledge are conceivably manifold: it could, for instance, be (1) provided as an attribute to end-users browsing documents, (2) used as metadata by search engines, (3) used in *knowledge audits* and *knowledge gap analyses*, enabling organizations to learn when knowledge in a particular area needs to be consolidated. It is, of course, also applicable in a more general information extraction sense: information that is extracted from text needs to have a certainty indicator attached to it.

A dimension other than certainty that has a clear impact on knowledge is precision. It would be interesting to evaluate the reliability of knowledge based on a combination of certainty and precision.

The annotated *World Bank* corpus will be made available for further research on the Web.

References

Mona T. Diab, Lori Levin, Teruko Mitamura, Owen Rambow, Vinodkumar Prabhakaram, and Weiwei Guo. 2009. Committed belief annotation and tagging. In *Proceedings of the Third Linguistic Annotation Workshop, ACL-IJCNLP*, pages 68–73, Suntec, Singapore, August. ACL and AFNLP.

Ken Hyland. 1998. *Hedging in Scientific Research Articles*. John Benjamins Publishing Company, Amsterdam/Philadelphia.

Halil Kilicoglu and Sabine Bergler. 2008. Recognizing speculative language in biomedical research articles: a linguistically motivated perspective. *BMC Bioinformatics*, 9.

George Lakoff. 1973. Hedges: A study in meaning criteria and the logic of fuzzy concepts. *Journal of Philosophical Logic*, 2:458–508.

Roser Morante and Walter Daelemans. 2009. Learning the scope of hedge cues in biomedical texts. In *Proceedings of the Workshop on BioNLP*, pages 28–36, Boulder, Colorado, June. Association for Computational Linguistics.

Philip V. Ogren. 2006. Knowtator: a protégé plug-in for annotated corpus construction. In *Proceedings of the 2006 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 273–275, Morristown, NJ, USA. Association for Computational Linguistics.

Dennis Reidsma and Rieks op den Akker. 2008. Exploiting 'subjective' annotations. In *HumanJudge '08: Proceedings of the Workshop on Human Judgements in Computational Linguistics*, pages 8–16, Morristown, NJ, USA. Association for Computational Linguistics.

Victoria L. Rubin, Elizabeth D. Liddy, and Noriko Kando. 2006. Certainty identification in texts: Categorization model and manual tagging results. In *Computing Affect and Attitutde in Text: Theory and Applications*. Springer.

The World Bank Group. 2010. Documents & Reports. `http://go.worldbank.org/3BU2Z3YZ40`, *Accessed May 13, 2010*.

Veronika Vincze, György Szaarvas, Richárd Farkas, György Móra, and János Csirik. 2008. The bioscope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinformatics*, 9.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39:165–210.

J. W. Wilbur, A. Rzhetsky, and H. Shatkay. 2006. New directions in biomedical text annotation: definitions, guidelines and corpus construction. *BMC Bioinformatics*, 7:356+, July.