Crowdsourced Hedge Term Disambiguation

Morgan Ulinski and Julia Hirschberg
Department of Computer Science
Columbia University
New York, NY, USA
{mulinski,julia}@cs.columbia.edu

Abstract
We address the issue of acquiring quality annotations of hedging words and phrases, linguistic phenomena in which words, sounds, or other constructions are used to express ambiguity or uncertainty. Due to the limited availability of existing corpora annotated for hedging, linguists and other language scientists have been constrained as to the extent they can study this phenomenon. In this paper, we introduce a new method of acquiring hedging annotations via crowdsourcing, based on reformulating the task of labeling hedges as a simple word sense disambiguation task. We also introduce a new hedging corpus we have constructed by applying this method, a collection of forum posts annotated using Amazon Mechanical Turk. We found that the crowdsourced judgments we obtained had an inter-annotator agreement of 92.89% (Fleiss’ Kappa=0.751) and, when comparing a subset of these annotations to an expert-annotated gold standard, an accuracy of 96.65%.

1 Introduction
Hedging refers to the use of words, sounds, or constructions that add ambiguity or uncertainty to spoken or written language. Hedging can indicate a speaker’s lack of commitment to what they are saying or an attempt to distance themselves from the proposition they are communicating. Identifying hedging behavior in conversational speech and text can also reveal information about social and power relations between conversants. Additionally, since hedging can be indicative of a lack of speaker commitment, identifying hedging is of interest to the information extraction community, to determine the extent to which statements have been believed by the writer or speaker.

A major challenge in identifying hedges is that many hedge words and phrases are ambiguous. For example, in (1a), appear is used as a hedge word, but not in (1b).

(1) a. The problem appears to be a bug in the software.
   b. A man suddenly appeared in the doorway.

Currently there are few corpora annotated for hedging, and these are available in a limited number of genres. In particular, there is currently no corpus of informal language annotated with hedge behavior. Acquiring expert annotations on text in other genres can be time consuming and may be cost prohibitive, which is an impediment to exploring how hedging can help with applications based on text in other genres. To address these issues, we have developed a method of acquiring hedge annotations through crowdsourcing, by framing the hedge identification task as a simple word sense disambiguation problem. In this paper, we describe this method and also our use of Amazon Mechanical Turk to construct a corpus of forum posts labeled with hedge information.

In Section 2, we discuss related work. In Section 3, we describe how we constructed our dictionary of hedge terms and created the hedge and non-hedge definitions for each. Section 4 describes the crowdsourcing task in more detail and discusses the resulting corpus. We conclude in Section 5.

2 Related Work
Currently, there is limited material available for studying hedging. The CoNLL-2010 shared task on learning to detect hedges (Farkas et al., 2010) used the BioScope corpus (Vincze et al., 2008) of biomedical abstracts and articles and a Wikipedia corpus annotated for “weasel words”. Because of the domain-specific nature of these corpora,
they can be difficult to apply to other text genres, such as social media or blogs. Additionally, the Wikipedia definition of a weasel word is slightly different than that of a hedge. Weasel words include language referring to personal opinions and subjectivity (e.g. excellent, best) in addition to uncertainty and lack of speaker commitment. Thus, it may be difficult to use the Wikipedia corpus to study hedging as a phenomenon that is distinct from subjectivity. Both the BioScope corpus and the Wikipedia corpus were annotated by experts and/or trained linguists; as with any annotation task, acquiring new expert-annotated data can be time- and cost-prohibitive. Our work differs from these in that we annotate a corpus of documents containing more informal language—a collection of forum posts. Additionally, rather than relying on the availability of trained linguists to annotate the corpus, our work explores how we can use crowdsourcing to obtain hedge annotations.

To facilitate annotation by non-experts, we frame the annotation task as a word sense disambiguation problem rather than asking directly about hedging. Note that there is a precedent for reformulating hedge detection in this way: as a follow-up to the CoNLL-2010 hedge classification task, Velldal (2011) described a new approach to classification in which hedge detection was viewed as a simple disambiguation task, restricted to words that have previously been observed as hedge cues. Velldal transformed the CoNLL data for the binary classification task by defining the dictionary of potential hedge terms as any tokens that appeared as hedge cues in the training data; all unlabeled instances of these terms were assumed to be non-hedge usages. A classifier trained using this approach was found to outperform the systems presented at CoNLL-2010, which relied on standard methods of token-by-token or sentence-level classification. Our work extends the word sense disambiguation approach to the problem of obtaining hedging annotations on new corpora.

Crowdsourcing has been successfully used in the past for collecting annotations for word sense disambiguation. Chklovski and Mihalcea (2002) had users select the WordNet sense that most closely matched the definition of a word as used in a given sentence. Likewise, Akkaya et al. (2010) used Amazon Mechanical Turk (AMT) to annotate Subjectivity Word Sense Disambiguation (SWSD), a coarse-grained word sense disam-

Table 1: List of (potential) hedge words and phrases

| Relational Hedges | Propositional Hedges |
|-------------------|----------------------|
| according to, appear, arguably, assume, believe, consider, could, doubt, estimate, expect, feel, find, guess, hear, I mean, I would say, imagine, impression, in my mind, in my opinion, in my understanding, in my view, know, likely, look like, looks like, may, maybe, might, my thinking, my understanding, necessarily, perhaps, possibly, presumably, probably, read, say, seem, seemingly, should, sound like, sounds like, speculate, suggest, suppose, sure, tend, think, understand, unlikely, unsure |
| a bit, a bunch, a couple, a few, a little, a whole bunch, about, allegedly, among others, and all that, and so forth, and so on, and suchlike, apparently, approximately, around, at least, basic, basically, completely, et cetera, etc, fair, fairly, for the most part, frequently, general, generally, in a way, in part, in some ways, kind of, kinda, largely, like, mainly, more or less, most, mostly, much, occasionally, often, partial, partially, partly, possible, practically, pretty, pretty much, probable, rarely, rather, really, relatively, rough, roughly, seldom, several, something or other, sort of, to a certain extent, to some extent, totally, usually, virtually |

Our work follows these examples by presenting hedging and non-hedging definitions and asking users to choose between them.

3 Hedging Dictionary

We compiled a dictionary of 117 potential hedge words and phrases. We began with the hedge terms identified during the CoNLL-2010 shared task (Farkas et al., 2010), along with synonyms of these terms. This list was further expanded and edited through consultation with the LDC and other linguists, to ensure representation of hedge terms from more informal text.

The full list of hedge words and phrases in our dictionary is shown in Table 1. This hedging dictionary is divided into relational and propositional hedges. As described in Prokofieva and Hirschberg (2014), relational hedges have to do
Table 2: Example definitions from our hedging dictionary

| Hedge Term | Hedge Definition | Non-Hedge Definition |
|------------|------------------|----------------------|
| about      | • almost; approximately ("There are about 10 million packages in transit right now.") | • on the subject of; concerning ("We need to talk about Mark.") |
|            | • located in a particular area ("He is about the house.") | • on the verge of ("He was about to leave.") |
| practically | • virtually; almost; nearly ("Their provisions were practically gone." "It has rained practically every day.") | • in a practical manner; realistically; sensibly ("Practically speaking, the plan is not very promising." "He purchased as many items as he could practically afford.") |
| suppose    | • to believe or assume as true ("It is generally supposed that his death was an accident.") | • to be expected or designed; to be required or permitted ("The machine is supposed to make noise." "I’m supposed to call if I’m going to be late.") |
|            | • to think or hold as an opinion ("I suppose the package will arrive next week.") | |
| think      | • to have an opinion, belief, or idea about someone or something ("I think it’s an important issue." "John doesn’t think he will win the election.") | • to use one’s mind actively to form ideas ("Think carefully before you begin." "I didn’t think of the solution in time.") |
|            | • to have as a plan or intention ("I thought that I would go.") | • to direct one’s mind toward something or someone ("I was thinking about you.") |

with the speaker’s relation to the propositional content, while propositional hedges are those that introduce uncertainty into the propositional content itself. The examples in (2) demonstrate relational and propositional hedges.

(2) a. I **think** the ball is blue.
    b. The ball is **sort of** blue.

In (2a), **think** is a relational hedge. In (2b), **sort of** is a propositional hedge.

For each hedge term in our dictionary, we created definitions for the hedging and non-hedging usages of the term, including examples for each case. We attempted to keep these definitions as simple as possible while still providing enough direction for workers completing the AMT task. These definitions were revised as we tested the AMT task with real-world users and received feedback pointing out ambiguities or other problems with the definitions. We did find that some words were too complicated or that the differences in senses was too nuanced to reduce definitions to short hedge and non-hedge definitions: in particular, **hear, read,** and **say** were identified as such. For example, the sentences in (3) differ only slightly, but **hear** is being used a hedge in the first and not in the second:

(3) a. I **heard** that there was an arrest.
    b. I **heard** about the arrest.

For these words, it might be more effective to develop a separate AMT task that provides a more comprehensive set of definitions and examples rather than trying to reduce them to a simple binary choice. Another option would be to ask AMT workers more directly about how the speaker is using a term: e.g. whether the usage reflects uncertainty or lack of commitment to a proposition. Table 2 shows some examples of hedging and non-hedging definitions. The complete dictionary of hedge terms, definitions, and examples is available from the authors upon request. Note that for 34 entries in our dictionary, the **non-hedge** definition is simply “Other”. These are cases where the word or phrase is generally unambiguous except for extremely rare instances (generally, typos...
4. It seems a bit silly now but I was fairly young when we met and this seemed so important to me.

Is the meaning of the word fairly closer to:

- justly or honestly, impartially ("The judge decided fairly.")
- to quite a high degree ("I was fairly certain she had nothing to do with it.")
- somewhat ("This judgment passed down fairly recently.")

Figure 2: Example of AMT word sense disambiguation task
The agreement varied depending on the hedge term. Figure 4 shows a scatterplot of the agreement percentage vs. how often each term is used as a hedge. As one might expect, the general trend shows that agreement is higher for terms that are almost always used as hedges (or as non-hedges) than for the more ambiguous terms.

To get a sense of the quality of the crowdsourced judgments, we annotated a subset of the corpus ourselves. This subset was constructed by randomly selecting two instances for each hedge term. Each instance received two judgments, one by each of the two authors of this paper. As one would expect, inter-annotator agreement was higher, 94.73% overall, with Cohen’s Kappa equal to 0.857. For most hedge terms, agreement was 100%; however, 11 hedge terms had an agreement of 50%. We adjudicated the questions for which we disagreed to create a single gold standard answer. We then compared our gold standard answers for this subset to the majority vote judgments obtained from AMT workers for the same questions. The crowdsourced majority vote judgment differed from the gold standard on only 7 questions, for an overall accuracy of 96.65%.

5 Summary

We have described a new method of using crowdsourcing to annotate a corpus with hedging information, by framing the hedge detection task as a word sense disambiguation problem. We have used this method to annotate a corpus of forum posts, which we hope to make generally available through the LDC. We have shown that annotations obtained using this method can in fact be very accurate; when comparing the crowdsourced judgments to an expert-annotated subset of the corpus, we obtained an accuracy of 96.65%.

Acknowledgments

This paper is based upon work supported by the DARPA DEFT program. The views expressed here are those of the author(s) and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

References

Cem Akkaya, Alexander Conrad, Janycz Wiebe, and Rada Mihalcea. 2010. Amazon Mechanical Turk for Subjectivity Word Sense Disambiguation. In Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon’s Mechanical Turk, pages 195–203, Los Angeles. Association for Computational Linguistics.

Timothy Chklovski and Rada Mihalcea. 2002. Building a Sense Tagged Corpus with Open Mind Word Expert. In Proceedings of the ACL-02 Workshop on Word Sense Disambiguation: Recent Successes and Future Directions, pages 116–122. Association for Computational Linguistics.

Richárd Farkas, Veronika Vincze, György Móra, János Csirik, and György Szarvas. 2010. The CoNLL-2010 Shared Task: Learning to Detect Hedges and their Scope in Natural Language Text. In Proceedings of the Fourteenth Conference on Computational Natural Language Learning, pages 1–12, Uppsala, Sweden. Association for Computational Linguistics.

Jennifer Garland, Stephanie Strassel, Safa Ismael, Zhiyi Song, and Haejoong Lee. 2012. Linguistic resources for genre-independent language technologies: User-generated content in BOLT. In @NLP Can u tag #user_generated_content ?? Via LrecConf.Org (NLP4UGC 2012).

Anna Prokofieva and Julia Hirschberg. 2014. Hedging and speaker commitment. In Proceedings of the 5th International Workshop on Emotion, Social Signals, Sentiment & Linked Open Data, pages 10–13, Reykjavik, Iceland.

Rion Snow, Brendan O’Connor, Daniel Jurafsky, and Andrew Ng. 2008. Cheap and Fast – But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks. In Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, pages 254–263, Honolulu, Hawaii. Association for Computational Linguistics.

Erik Velldal. 2011. Predicting speculation: A simple disambiguation approach to hedge detection in biomedical literature. Journal of Biomedical Semantics, 2(5):S7.

Veronika Vincze, György Szarvas, Richárd Farkas, György Móra, and János Csirik. 2008. The BioScope corpus: Biomedical texts annotated for uncertainty, negation and their scopes. BMC Bioinformatics, 9(11):S9.