Corpus Annotation as a Scientific Task

Donia Scott, Rossano Barone, Rob Koeling

Department of Informatics,
University of Sussex Brighton, UK.
{D.R.Scott, R.Barone, R.Koeling}@sussex.ac.uk

Abstract

Annotation studies in CL are generally unscientific: they are mostly not reproducible, make use of too few (and often non-independent) annotators and use guidelines that are often something of a moving target. Additionally, the notion of 'expert annotators' invariably means only that the annotators have linguistic training. While this can be acceptable in some special contexts, it is often far from ideal. This is particularly the case when subtle judgements are required or when, as increasingly, one is making use of corpora originating from technical texts that have been produced by, and intended to be consumed by, an audience of technical experts in the field. We outline a more rigorous approach to collecting human annotations, using as our example a study designed to capture judgements on the meaning of hedge words in medical records.

Keywords: annotation, hedges, electronic patient records

1. Introduction

Even when undertaken with the aid of one of the many excellent annotation toolkits available, human corpus annotation is a time-consuming and cognitively loaded task. It also typically requires a lot of training, with carefully-constructed guidelines and regular monitoring. Most human annotation tasks thus involve a small number of annotators, and although the standard is moving towards triple-annotated documents, many studies still include a number of single annotated documents (see, e.g., Thompson (2008)). This scenario (even with triple-annotations) is not ideal in that it is unlikely to lead to robust models[1], especially in cases where the annotation task requires subtle judgements for which there is no established gold standard. Such cases instead require careful analysis of a representative set of language samples in a highly controlled laboratory study involving a substantial number of annotators. This is the purview of experimental psycholinguistics.

We describe here our use of standard techniques from experimental psycholinguistics (and more generally, experimental psychology) for gathering human judgements of linguistic data from an established corpus. We ground our description in an annotation study we have conducted on the use of linguistic hedges in medical communication, as a preliminary to applying machine-learning techniques for information extraction of medical decisions/opinions in a large corpus of electronic patient records.

2. Vagueness in medical communication

> "Medical vagueness refers to the inherent and irreducible uncertainty that occurs when clinical knowledge is applied to specific patients"(Emanuel and Emanuel, 1989)

This vagueness is mostly expressed in medical documents through the use of linguistic hedges – i.e., modifiers such as *"possibly* <diabetes>", *"probably* <diabetes>", *"consistent with* <diabetes>" etc. (see, e.g., Hyland (2006)). Consider for example these fairly typical excerpts taken from radiology reports:

> *Possible early pneumonia involving the lingual and possibly the right middle lobe.*
>
> *Mild perihilar bronchial wall thickening may represent either viral infection or reactive airways disease.*
>
> *Subtle ill defined opacity in the left lower lobe could represent pneumonia in the appropriate clinical setting.*
>
> *Left posterior lung base opacity which appears somewhat homogeneous which is somewhat atypical for consolidation.*

The issue for us, is to understand the level of (un)certainty of the state of affairs that these various hedges are intended to convey, especially in a medical setting. More specifically, faced with the task of extracting information about diagnoses, symptoms etc. from electronic patient records, how should we treat information that is couched in speculative language? To achieve a robust model of the mappings between specific hedges and levels of (un)certainty, we need to attend to a number of methodological factors.

3. Methodological requirements

3.1. Variability between and within annotators

Studies in non-medical domains have suggested that the assignment of certainty levels to hedges is highly variable (Budescu and Wallsten, 1985). Variability may occur for a range of reasons including context effects of the sentence on hedge interpretation, effects of prior beliefs about the hedge proposition and effects on accuracy resulting from differing levels of conscientiousness/fatigue that may occur over the course of the annotation episode. The challenge is

[1] See Krippendorff (2004) for an extended discussion of this point.)

therefore to design a study that will control for unwanted variability and allow us to derive statistical idealisations that account for such effects.

3.2. Method of data collection

For good quality annotations theres an obvious requirement that the annotators themselves are well qualified to make the judgements. In a case such as ours, we need to make sure that the annotators are very experienced in the language of medical records as regular authors and/or readers of such documents. Additionally, the study requires a large sample of such time-committed individuals who will be prepared to engage fully in the task and to carry it out to completion. It is difficult to get professional medics to devote the time required for annotation. The challenge is therefore to employ an alternative method of data collection that will be effective in recruiting a large sample of appropriately trained annotators whilst providing reliable data.

3.3. Efficacy of the annotation task

In order to ensure that all annotators are applying the same criteria for any given judgement, the standard procedure in corpus linguistics is to develop annotation guidelines that provide clear criteria for each tag in the annotation set. This of course requires some a priori understanding of meaning of the elements in the tagset (e.g., named entities, conference, parts of speech, diagnoses/symptoms etc.). Another important challenge for any annotation task is to make it efficient in terms of both the required time and cognitive effort needed to perform the task, and also its capacity to be correctly understood and followed. Ideally, the task should minimise opportunities for errors and unconscientious judgments, thus improving overall data quality. There are a range of task-design features that are critical to meeting these challenges, including ensuring that the scheme and instructions are simple, intuitive and unambiguous, and presenting the test materials in ways that facilitate correct comprehension.

3.4. Assessing the validity of annotations

An important requirement for determining data quality is to eliminate data resulting from participants who do not adhere to the task instructions. A critical methodological challenge is therefore to design methods for identifying invalid data resulting from non-compliance of task instructions.

4. Solution methodology

We designed our annotation study as a web-based psycholinguistic experiment that required a forced-choice response to hedged sentences taken from a corpus of medical records. Figure 1 shows a snapshot of what participants saw. With each response they make, subjects in the study are effectively annotating the hedge that appears in the sentence under consideration.

4.1. Corpus

We made use of the BioScope corpus of electronic medical records, which contains 855 sentences already tagged for speculative language (i.e., hedges and their scope) (Vincze et al., 2008). We manually cleaned up the corpus to remove obvious annotation errors and to eliminate ambiguous or semantically complex cases, involving e.g., alternatives, coordination, or multiple hedges. This left us with a pool of 313 sentences. The distribution of hedges in the corpus followed the typical Zipf profile: ranging from 108 tokens of "may" (e.g., *"Mild perihilar bronchial wall thickening may represent reactive airways disease."*) to many cases for which there were only a few, or even only one, token. Of these we selected the top most frequent hedges types with the cutoff point being ones for which there were six of more tokens; this gave us 18 types of hedges. Finally, we randomly selected six tokens of each type[2]. This provided 108 (6 x 18) hedged sentences for our study.

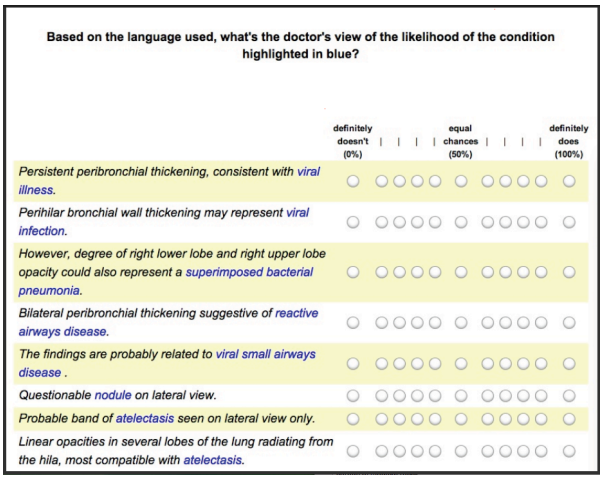

Figure 1: Snapshot of an example of participants' view of the study

For each of the selected sentences, we highlighted the scope of the hedge (in blue, as shown in Figure 1). The highlighting feature was aimed at facilitating task consistency by ensuring that the object (i.e., scope) of the hedge was clearly signalled. Highlighting only the scope rather than the hedge itself was intended to make it more difficult for participants to simply work through the sentences without reading them fully, making decisions by looking at the hedge word(s) only.

4.2. Variability between and within annotators

To address this issue, we chose to collect data from a large sample of subjects/annotators and, from each a sample of responses to the same hedge. This approach to providing estimates of the central tendency of values and dispersion (as normally employed in psychology experiments) will allow us to identify and account for the variability effects described above, and provide an approximate measure of the nature of such variability within the population.

4.3. Annotation task

Tags were selected through an answer to the question:

> *Based on the language used, what's the doctor's view of the likelihood of the condition highlighted in blue?*

[2]Where there were only six tokens, we chose all six.

Many linguistic studies on hedges employ a classification schema that involves a probability scale attached to lexical labels. The problem with this is that on the one hand, classifying with the verbal labels forces readers to map one hedge onto to yet another, which obviously introduces unwanted circularity; on the other hand, classifying according to probability introduces complex problems of its own (see Clark (1990) for a discussion). To avoid this, we chose instead to employ a likelihood scale (as shown in Figure 1) that comprises 11 points that can be mapped onto values from 0% to 100% with 10% intervals. Although 0% and 100% may be considered unrealistic commitments, the design of the scale, including the minimum and maximum slots, provides a familiar metric frame of reference for assigning likelihood values that can be treated as interval data for data analysis.

4.4. Method of data collection

The study used SurveyMonkey, a web based survey tool, to collect annotation data. We acquired annotators through targeted emails sent to professional newsgroups in medicine and biomedicine and to colleagues in medical schools. To control for possible effects of professional experience, we gathered information from participants on their medical training. We also collected information on their native language; given the subtlety of the judgements required, and the well-known differences of some distinctions between British and American English, this allowed us to control and thus account for effects of language experience.

4.5. Annotation procedure

For the core part of the study, participants were required to work their way through six pages of annotation materials (such as that shown in Figure 1). Each contained one token of each of the 18 hedges, plus a dummy sentence. Participants were not able to advance to the next page unless they had responded to all sentences. To avoid practice/fatigue effects and the adoption of strategies that involve reusing (from memory or the survey) a previously given response to the same hedge, the order of presentation of sentences within each page was randomised per participant, and it was not possible for participants to look back at pages they had already completed.

Prior to embarking on the main study, participants completed requests for profile information, and went through a short practice session. The full task took about 20 minutes to complete.

4.6. Eliminating invalid annotations

The dummy sentences were included to aid identification of problematic data. They each involved a (different) hedge whose likelihood values were judged by us to be at one of the extreme ends of the likelihood scale (e.g., *"It is certain that the patient has pneumonia"*). As the prejudged likelihood ratings of the majority of test hedges were distributed around the centre rather than the extreme ends of the likelihood scale, the probability of an arbitrary choice conforming with the typical distribution of values by chance was thus minimised.

5. Validity of the methodology

The focus of this paper is to describe a scientifically rigorous method for collecting human judgements/annotations of linguistic data. The method we describe here ensures that the study is reproducible (Krippendorff, 2004):

- the task is intuitive;
- the instructions are clear and the opportunities for scoping ambiguities are close to zero;
- all participants/coders work independently using exactly the same instructions;
- the choice of coders is clearly specified, and there are opportunities to examine the differences between different categories of coders; clear criteria are formulated for identifying errant coders;
- the coding materials are a representative sample of a 'clean' subset of the corpus, thus reducing the opportunities for context effects and/or ambiguities.
- the reliability of responses is enhanced by the use of repeated measures; many (and the same) judgements were collected from each participant for each hedge type, and all participants judged the full set of materials;
- the opportunities for practice effects are reduced (and controlled), since the materials are presented to the participant coders each in a different random order.

Turning to the results of the study: we collected 13,176 annotations of the 18 hedges from 122 annotators, at a total cost of £100.

Our assessment of the within-[3] and between-annotator[4] consistency (as indexed by measures of variability/dispersion) suggests that nearly all participants correctly adhered to the task instruction; only a small number had to be eliminated (3 in 122). Of these, two were eliminated on basis of atypical responses on the dummy sentences and also giving the same responses to all the statements. Although the particular design of the dummy sentences was a useful feature, it was not a sensitive to all problem data. For example, one respondent consistently made a likelihood judgement of "definitely does/100%" to all *"no evidence"* hedge statements (e.g., *"No evidence of acute cardiopulmonary disease"*); this suggests that he/she may have confused the goal of determining *objective*

[3]The *within* consistency value is computed by taking the SD of the ratings for a single participant to a set of hedge instances belonging to a single hedge type. Let's call this a *w* value for a hedge type. The *within* values reported in this paper is the average for all *w* values across all hedge types and across all participants.

[4]The *between* consistency value is computed by deriving the mean ratings for a single participant to a set of hedge instances belonging to a single hedge type. The standard deviations of these means between all participants for a given hedge types are then derived for each hedge type. Hence you will have list of *between* subject SDs for each hedge type. The average *between* consistency value in the paper is the mean of all the SDs for a hedge type.

likelihood with *certainty of belief*. Hence in addition to the dummy manipulations, the computation of deviant test scores also provided useful grounds for identifying potentially problematic data.

Focussing our attention on only those participants who are native speakers of British or American English, the results show that variation, as measured by standard deviation, is low both within (SD=7.79%) and between (SD=8.95%) annotators. The low standard deviations indicate that the results for each hedge are clustered closely around its mean. The within-participant result shows that even though the expressions being examined are inherently vague, participants' judgements are typically (about 68%, assuming a normal distribution) within a point on either side of their mean judgement score (on the scale shown in Figure 1). The low between-participant result is particularly telling: it shows that participants' judgements tend to all fall very much within the same range on the scale.

A 2 x 2 x 18 mixed Analysis of Variance (ANOVA) was conducted to investigate effects of training group, language group and hedge types on within-subject consistency with repeated measure on the later factor. The ANOVA with the Greenhouse-Geisser correction revealed that the main effect of hedge type was highly statistically significant ($F(8.47, 753.97) = 8.22$, $P>0.0005$). The main effects of training group and language group were not statistically significant and neither were any of the interactions between the three factors. Average SDs for within-subject consistency ranged from 5.27% for the *'no evidence'* hedge type to 11.00% for the *'questionable'* hedge type. These results are consistent with the hypothesis that hedge types differ in terms of the semantic vagueness of the likelihood they convey and suggest that the methodology employed could be used to provide a statistical measure of such vagueness.

The data allowed differences between hedge types for between-subject consistency to be identified. Standard deviations for between-subject consistency ranged from 6.72% for the *'most likely'* hedge type to 14.32% for the *'questionable'* hedge type. Such data provides detailed measures of agreement between participant for particular hedge types contingent on the sampling methodology employed in the study.

The result of the data suggest that the methodology supports higher-order classification of hedge types. For example, a high degree of similarity in average likelihood ratings for particular groups of hedge types were observed for hedge types within the following groups [*'possibly', 'possible', 'may'*], [*'suggests', 'likely', 'probably', 'probable'*] and [*'consistent with', 'most consistent with', 'most likely', 'most compatible with'*] differing amongst each other only by approximately 1% or less. Such grouping were readily accessible in graphical representations of the data.

The observed variability is consistent with the proposal that participants responded to each hedge-sentence on it own terms rather than employ response-reuse strategies, and that there was a high degree of consistency in their judgements/annotations; the design features of the study are likely to have played a significant role in helping to identify and eliminate sources of dirty data. The observed variability also suggests that the acquired model of the meanings of the selected hedges in a medical setting will be robust.

6. Conclusions

Our approach provides a clear and scientific method for gathering annotations on a corpus. It also provides a new alternative to crowd-sourcing (e.g., via Amazon Mechanical Turk) as a means of gathering large numbers of annotations at low cost. It has been suggested that crowd-sourcing of annotations leads to no meaningful loss in data quality compared to that obtained from 'expert' annotators — at least for the tasks studied (affect recognition, word similarity, recognising textual entailment, event temporal ordering and word sense ambiguation) (Snow et al., 2008). The term 'expert' here applies to *linguistic* experience.

Our data suggest that the another kind of 'expertise' of annotators can be a critical factor, at least when subtle judgements about a domain are required: *expertise in the sublanguage of the domain*. Our results show a (small but) significant main effect of the domain expertise of our annotators (2-way ANOVA, $F(1, 89)=4.89$, $p<.05$), with medically-trained annotators consistently judging hedges as expressing a greater likelihood — in other words, when a doctor reads, for example, that in the view of a colleague a given patient *"probably has diabetes"*, he or she will generally interpret the likelihood of that patient having diabetes as greater than, say, the patient, would (assuming of course that the patient is not also a medical professional). The imperative to seek medical experts to annotate medical language is an important finding of this study; while it has been recognised in another recent study of electronic medical records by Roberts et al. (2007), whose annotators "include" clinicians, bioinformaticians and medical students, it is a point that is often overlooked. General linguistic expertise may be useful for many annotation tasks, but it is not always enough and indeed, it is not always relevant.

7. Acknowledgements

We wish to thank Jackie Cassell and Helen Smith, our clinical colleagues at the Brighton and Sussex Medical School, for their contribution to this study. The work was supported by the Wellcome Trust [086105/Z/08/Z] as part of the Patient Records Enhancement Project (PREP). The authors were independent from the funder and sponsor, who had no role in conduct, analysis or the decision to publish.

8. References

D. Budescu and T. Wallsten. 1985. Consistency in interpetation of probabilistic phrases. *Organizational Behavior and Human Decision Processes*, 36:39 – 405.

D.A. Clark. 1990. Verbal uncertainty expressions: A critical review of two decades of research. *Current Psychology: Research and Reviews*, 9(3):203 – 235.

L. Emanuel and E.J. Emanuel. 1989. The medical directive: A new comprehensive advance care document. *Journal of the American Medical Association*, 261(22):3288 – 3293.

K. Hyland. 2006. Medical discourse: hedges. In K. Brown, editor, *Encyclopedia of Language and Linguistics*, page 694697. Elsevier, Oxford, 2 edition.

K. Krippendorff. 2004. *Content Analysis: An Introduction to Its Methodology*. Sage, Thousand Oaks, CA, 2nd edition.

A. Roberts, R. Gaizauskas, M. Hepple, N. Davis, G. Demetriou, Y. Guo, J. Kola, I. Roberts, A. Setzer, A. Tapuria, and B. Wheeldin. 2007. The CLEF corpus: Semantic annotation of clinical text. In *Proceedings of the Annual Symposium of the American Medical Informatics Association (AMIA, 2007)*, pages 625 – 629.

R. Snow, B. OConnor, D. Jurafsky, and A.Y. 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.

P. Thompson. 2008. Building a bio-event annotated corpus for the acquisition of semantic frames from biomedical corpora. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may. European Language Resources Association (ELRA).

V. Vincze, G. Szarvas, R. Farkas, G. Móra, and J. Csirik. 2008. The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinformatics*, 9(Suppl 11):S9.