A Study of Inter-Annotator Agreement for Opinion Retrieval

Adam Bermingham and Alan F. Smeaton
CLARITY: Centre for Sensor Web Technologies
Dublin City University
Dublin, Ireland.
{abermingham,asmeaton}@computing.dcu.ie

ABSTRACT

Evaluation of sentiment analysis, like large-scale IR evaluation, relies on the accuracy of human assessors to create judgments. Subjectivity in judgments is a problem for relevance assessment and even more so in the case of sentiment annotations. In this study we examine the degree to which assessors agree upon sentence-level sentiment annotation. We show that inter-assessor agreement is not contingent on document length or frequency of sentiment but correlates positively with automated opinion retrieval performance. We also examine the individual annotation categories to determine which categories pose most difficulty for annotators.

Categories and Subject Descriptors

H.3.4 [**Information Retrieval**]: Systems and Software Performance evaluation (efficiency and effectiveness)

General Terms

Experimentation, Measurement, Human Factors

1. INTRODUCTION

With the abundance of user-generated content on the Internet in recent years, there has been much effort to model the sentiment in online texts. Annotated documents are necessary to evaluate systems designed to automatically classify, rank or score documents with respect to opinion. In some domains, such as film reviews, a sentiment polarity score is often readily available as users annotate their documents with a quantified summary e.g. *4 out of 5 stars*. In other domains an author annotation is not available and we rely on human assessors to create annotations or judgements. There are a number of subjective variables associated specifically with opinion annotation which affect agreement including domain expertise, personal opinion, ambiguity of language, and context of interpretation. One other issue is granularity of sentiment and previous annotation efforts have varied from the document level [3] to sentence- and sub-sentence-levels [5]. There have also been efforts at multi-lingual sentence-level opinion annotation which have yielded moderately high rates of agreement for Japanese and Chinese but low agreement for English texts [4].

SIGIR'09, July 19–23, 2009, Boston, Massachusetts, USA.
ACM 978-1-60558-483-6/09/07.

Using documents from the *Blogs06* corpus used at the TREC Blog Track [3], we asked participants to identify opinion at sentence-level. We then measure sentence-level inter-annotator reliability for all sentences and repeat this for each document and topic. Extrapolating annotations to the document-level, we then draw comparisons between sentence-level and document-level agreement. Finally, we convert the annotations to binary judgements for each annotation class to allow per-class analysis.

2. EXPERIMENTAL SETUP

Our 15 participants were postgraduate students and postdoctoral researchers, 5 of whom have worked in sentiment analysis and 13 of whom were native English speakers. 15 topics were selected out of the 150 topics used in the Blog Track at TREC 2008 based on median TREC participant performance per topic, evenly distributed from low-performing topics to high-performing topics. A pool of documents was selected from our own baseline TREC run [1], up to a maximum of 8 documents per topic. All of the documents selected were judged by the TREC relevance assessments to contain opinion on the topic and consisted of plain text blog entries extracted from HTML and passed through the noise removal portion of our TREC system.

In the annotation process, participants were presented with a series of 30 topic/document pairings and asked to annotate the sentences in each document as one of five categories: "*non-relevant*", "*relevant*" (relevant and no opinion), "*positive*", "*negative*", "*mixed*". When a document is initially presented to a participant, all of the sentences are annotated as *non-relevant*. After completing the sentence-level annotation, participants were then asked to rate the document for negative opinion from 1 ("*no negative topic-directed opinion*") to 5 ("*very obvious and intense negative topic-directed opinion*") and similarly for positive opinion.

In total, 115 documents were judged by an average of 3.6 annotators yielding 26,375 sentence annotations.

3. RESULTS AND EVALUATION

We use Krippendorff's *alpha*[2] for measuring inter-annotator agreement. This is a robust statistic which takes into account the probability that observed variability is due to chance and does not require that each annotator annotate each document. α for sentence-level annotation with respect to the 5 classes in Section 2 is 0.4219. This indicates a significant agreement between annotators but is less than the level recommended by Krippendorff for reliable data ($\alpha = 0.8$) or for tentative reliability ($\alpha = 0.667$).

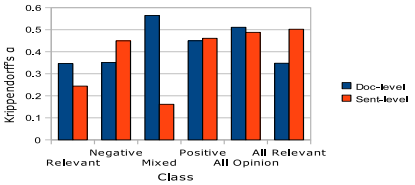

Figure 1: α for Binary Judgements

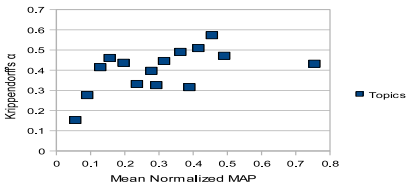

Figure 2: α and mean TREC MAP ($\rho = 0.53$,$\tau = 0.41$)

If we examine α for each of the 115 documents, we see little correlation between α and the number of sentences per document (Pearson's $\rho = -0.123$, Kendall's $\tau = -0.13$) or between α and the proportion of sentences annotated as containing opinion ($\rho = -0.045$, $\tau = -0.015$). This indicates that the consistency between annotators is not dependent upon the proportion of sentiment-bearing sentences or the overall length of the document.

Calculating α for each of the 15 topics, we see a significant positive correlation between the retrieval performance of the topics at TREC and α for each topic ($\rho = 0.53$, $\tau = 0.41$). This reflects the increased ambiguity and obscurity among the low-performing topics which hampers both automated opinion retrieval and manual annotation efforts similarly. A ranking of the 15 topics by α demonstrates no discernable pattern in terms of topic nature.

In order to simulate document-level annotations, we extrapolate document-level annotations from sentence-level annotations. For example, a document containing positive sentences but no mixed or negative sentences would be considered a positive document. Agreement for these document-level annotations is slightly higher than for sentence-level ($\alpha = 0.4461$) suggesting that although annotators may differ in their reasons for document annotation, they converge a small amount at document-level. It should be noted that the simulated document-level annotations are not necessarily the same as would be obtained had the annotators been explicitly asked to annotate at the document-level.

To look at the individual classes more closely we map the 5-way annotations to binary judgements for each of the classes (Figure 1). The most striking difference in agreement between document and sentence-level annotation is for the *mixed* class. Agreement for this class is highest at the document-level and lowest at the sentence-level. It is also worth noting that there is much less agreement for negativity than positivity at document-level and that agreement is very low for the *relevant* class, particularly at the sentence-level. If we look at an additional aggregate class, *All Relevant*, there is a surprisingly low α for extrapolated document-level relevance. This possibly reflects the fact that 11% of documents were annotated as non-relevant, despite none of them being judged that way at TREC.

Finally, we determine a binary opinion judgement from the two document-level 5 point scales. A document is defined as opinionated if either of the scales record a value greater than 1 for that document. If each of these opinion judgements is compared with its corresponding opinion annotation extrapolated from sentence annotations, we see a very high level of agreement ($\alpha = 0.8263$). This shows consistency between each annotator's sentence and document-level annotations.

4. CONCLUSIONS AND DISCUSSION

We have found that sentence-level sentiment annotation yields a moderate level of inter-annotator agreement and that this is independent of the nature of the sentiment, specifically the frequency of the sentiment and document length. We suggest that the 5 classifications used here (and in TREC) are not ideal categories for sentiment annotation. In particular, the *mixed* category shows very low agreement at sentence-level. At document-level there is high agreement but only due to the broad definition of *mixed* as a document containing both positive and negative opinions. This does not necessarily reflect the overriding sentiment in a document as both sides of a discussion are frequently cited in distinctly polarised documents, yielding an artificially high proportion of *mixed* documents.

Annotators reported frequently feeling uneasy about their judgements, particularly where domain or background knowledge was required. For this reason we suggest an *indeterminate* class which they are encouraged to use when they are not confident about their annotation. We would also like to examine the task description and annotator training more closely to see what effect it may have on agreement.

Finally, we show an increase in agreement can be achieved by simulating document-level judgements, suggesting that sentence-level annotation is too granular. For future work we would like to compare annotation at the document, paragraph, sentence and passage levels in an effort to identify the most appropriate sentiment granularity, both for annotation and automated opinion retrieval.

Acknowledgments

This work is supported by Science Foundation Ireland under grant 07/CE/I1147.

5. REFERENCES

[1] A. Bermingham, A. Smeaton, J. Foster, and D. Hogan. DCU at the TREC 2008 Blog Track. In *The Seventeenth Text REtrieval Conference (TREC 2008) Proc.*, 2008.

[2] A. F. Hayes and K. Krippendorff. Answering the call for a standard reliability measure for coding data. In *Communication Methods and Measures*, 2007.

[3] I. Ounis, C. MacDonald, and I. Soboroff. Overview of the TREC-2008 Blog Track. In *The Text REtrieval Conference (TREC 2008) Proc.* NIST, 2008.

[4] Y. Seki, D. K. Evans, L. Ku, L. Sun, H. Chen, and N. Kando. Overview of multilingual opinion analysis task at NTCIR-7. 2008.

[5] J. Wiebe, T. Wilson, and C. Cardie. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 1(2):0, 2005.