Carnegie Mellon University
Research Showcase @ CMU

Computer Science Department | School of Computer Science

5-2008

Vox Populi Annotation: Measuring Intensity of Ideological Perspectives by Aggregating Group Judgments

Wei-Hao Lin
Carnegie Mellon University

Alexander Hauptmann
Carnegie Mellon University, alex@cs.cmu.edu

Follow this and additional works at: http://repository.cmu.edu/compsci

Published In
Proceedings of the Sixth Language Resources and Evaluation Conference (LREC), .

This Conference Proceeding is brought to you for free and open access by the School of Computer Science at Research Showcase @ CMU. It has been accepted for inclusion in Computer Science Department by an authorized administrator of Research Showcase @ CMU. For more information, please contact research-showcase@andrew.cmu.edu.

Published in the Proceedings of the Sixth Language Resources and Evaluationa Conference (LREC), 2008.

Vox Populi Annotation: Measuring Intensity of Ideological Perspectives by Aggregating Group Judgments

Wei-Hao Lin and Alexander Hauptmann

Language Technologies Institute
School of Computer Science
Carnegie Mellon University
{whlin,alex}@cs.cmu.edu

Abstract

Polarizing discussions about political and social issues are common in mass media. Annotations on the degree to which a sentence expresses an ideological perspective can be valuable for evaluating computer programs that can automatically identify strongly biased sentences, but such annotations remain scarce. We annotated the intensity of ideological perspectives expressed in 250 sentences by aggregating judgments from 18 annotators. We proposed methods of determining the number of annotators and assessing reliability, and showed the the sentence-level annotations on ideological perspectives were reliable across different annotator groups.

1. Introduction

Polarizing discussions about political and social issues commonly occur in broadcast news, newspapers and blogs. We are interested in how contrasting ideological perspectives are expressed in spoken and written text. By *perspective*, we mean a point of view held by a group of people sharing similar cultural, social, and political beliefs. For example, in the discussions of the United States politics, the Democrats and Republicans are two major ideological perspectives.

In this paper we focus on annotating the intensity of ideological perspectives expressed at the sentence level. Annotations on the ideological perspectives at the document level are available (e.g., (Lin et al., 2006)); annotations at the sentence level, however, remain scarce. Not all sentences in a biased document express the overall ideological perspective to the same degree, and manual annotations are needed. For example, in the context of the Israeli-Palestinian conflict, Example 1 and Example 2 were written from the Israeli and Palestinian perspectives, respectively:

(1) The inadvertent killing by Israeli forces of Palestinian civilians – usually in the course of shooting at Palestinian terrorists – is considered no different at the moral and ethical level than the deliberate targeting of Israeli civilians by Palestinian suicide bombers.

(2) In the first weeks of the Intifada, for example, Palestinian public protests and civilian demonstrations were answered brutally by Israel, which killed tens of unarmed protesters.

Example 3, however, introduces an issue's general background and expresses less distinctly an ideological perspective:

(3) The Rhodes agreements of 1949 set them as the ceasefire lines between Israel and the Arab states.

Annotations on the sentence-level intensity of ideological perspectives will be very valuable for developing linguistic theories of ideological perspectives. The annotation will also be vital for evaluating computer programs that automatically extracted strongly biased sentences.

Annotating intensity of ideological perspectives, however, is challenging. The common practice for annotation intensity is to quantize intensity into discrete categories. For example, we could potentially allocate three categories (Strong, Medium, and Weak) for the Palestinian perspective and the Israeli perspective, plus one Neutral category, resulting on a total of seven categories. However, training annotators to agree on each of seven categories is not trivial. Instead, we ask annotators to make a simple binary decision: is the sentence more likely to be written from the Israeli perspective or the Palestinian perspective? We then aggregate binary decisions over a large number of annotators. While individual annotators have different thresholds on intensity of ideological perspectives, a sentence expressing strongly a Palestinian perspective will be likely to be labeled as Palestinian perspective by most annotators. On the other hand, a sentence expressing weakly a Palestinian perspective will have a mixed of the Israeli and Palestinian annotations.

- In this paper we annotated the intensity of ideological perspectives expressed in 250 sentences extracted from the web articles on the Israeli-Palestinian conflict (Section 3.1.).
- We quantitatively measured intensity of ideological perspectives by aggregating binary judgments from a group of annotators (Section 2.). Intuitively, strongly one-sided sentences would be more likely to be labeled consistently by a majority of annotators, while a neutral sentence would be equally likely to be labeled as displaying either perspective. We call this annotation method the Vox Populi Annotation.
- We want to ensure the reliability of Vox Populi Annotation method. How many annotators do we need to reliably estimate the intensity of the ideological perspective expressed in a sentence (Section 2.1.)? Are these intensity measures consistent across different groups of annotators (Section 2.2.)?

2. Vox Populi Annotation

We propose to quantitatively measure the degree to which a sentence expresses an ideological perspective by aggregating group judgments. We ask a group of annotators to make a forced binary choice on a sentence's ideological perspective, coded as 0 for Perspective A and 1 for the contrasting perspective B. A sentence's intensity is estimated to be the average of group judgments, ranging between 0 and 1. The larger the average value, the more intensely a sentence expresses Perspective B (and the more mildly the sentence expresses Perspective A). We call this annotation method as Vox Populi Annotation, and call the measure as Vox Populi Intensity.

The Vox Populi Annotation method is easy to implement. To annotate a sentence's intensity of expressing a particular ideological perspective, Vox Populi Annotation instructions can be as simple as "Which side do you think the sentence was written from?". Compared with most annotation studies, Vox Populi Annotation requires very little annotator training. However, are these intensity measures using the Vox Populi Annotation method *reliable*?

The Kappa statistic (Carletta, 1996; Artstein and Poesio, 2005b) cannot adequately assess the reliability of Vox Populi Annotations because annotators are not expected to agree on the same sentence at all. Contrary to most annotation studies, the Vox Populi Annotation method expects a large number of annotators to agree *collectively*, not on an individual basis. A sentence of intensity 0.75 is expected to have a quarter of n annotators who disagree with the other three quarters. If the Vox Populi Annotation method is indeed reliable , we will still expect to see considerable disagreements when the same sentence is labeled by a new group of n annotators, but the proportion should be close to 0.75.

Similar to the chance-corrected kappa statistic, we assess the reliability of Vox Populi Annotations by considering how much observed annotations can be attributed to random guessing.

- The Vox Populi Annotation method estimates a sentence's ideological intensity by aggregating group judgments, but how many annotators are needed to make reliable measurement? In Section 2.1. we quantify the exact relationship between the number of annotators and the desired reliability.
- After a group of annotators label a set of sentences, how do we assess whether these Vox Populi Intensities are random guesses? In Section 2.2. we propose a method of assessing the reliability of the Vox Populi Annotation method on a collection of sentences.

2.1. Number of Annotators

How many annotators do we need to be confident that Vox Populi Intensity is not random guessing? We see this question as a statistical testing problem, where the null hypothesis is $\mu = 0.5$, and the alternative hypothesis is $\mu \neq 0.5$, where μ is the mean of the intensity of a ideological perspective expressed in a sentence. The Vox Populi Annotation method requires annotators to make forced binary choices for each sentence, and each choice is like flipping a coin, i.e., a Bernoulli experiment.

We choose the exact Binomial test (Conover, 1971) to test the above hypothesis. The test procedure's power depends on two factors: the number of annotators and a sentence's ideological intensity (i.e., μ). There is a trade-off between two factors. If a sentence is extremely one-sided (i.e., μ is very close to 0 or 1), we do not need many annotators to reject the null hypothesis that a sentence is randomly annotated. However, more annotators are needed if a sentence mildly expresses an ideological perspective (i.e., μ is close to 0.5). Our confidence on the statistical testing procedure can be expressed as the p-value $p(x)$ of the exact binomial test, defined as follows:

$$p(x) = \sum_{i=0}^{x} \binom{n}{i} 0.5^n + \sum_{i=n-y}^{n} \binom{n}{i} 0.5^n,$$

where x is the number of annotators labeling a sentence as a particular perspective, n is the total number of annotators, and y is the number of integers between $\lceil n/2 \rceil$ and n whose binomial density under the null hypothesis is less than the density at x[1].

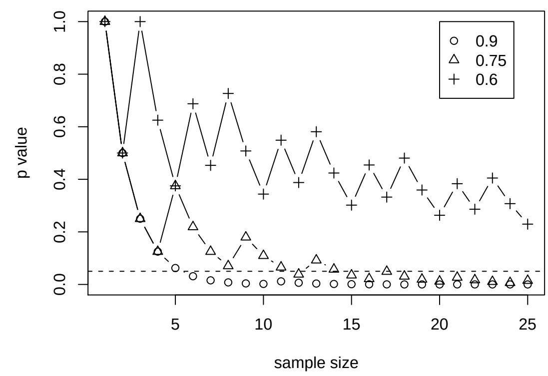

Figure 1: P-value decreases as an annotator group's size (sample size) increases. The horizontal dashed line is p-value 0.01. Three curves represent different Vox Populi Intensities. The curves zigzag due to a binomial distribution's discreteness.

We plotted the exact relationship between the number of annotators and and p-value for sentences of different Vox Populi Intensities in Figure 1. If by confident we mean p-value is less than 0.01 (the dash line), a sentence of intensity 0.9 (or 0.1 due to the symmetry of the binomial distribution under the null hypothesis) requires six or more annotators (the x axis) to reject the hypothesis that the sentence is randomly annotated. A sentence of intensity 0.75 (or 0.25) needs more than 18 annotators to reject the null hypothesis. A sentence of intensity 0.6 requires more than 100 annotators (not shown in Figure 1). Generally, the more annotators, the more confident we are that Vox Populi Intensity is not random; the more intensely a sentence expresses an

[1]We only list the case for $x < n/2$ and omit the case for $x \geq n/2$ because the two cases are very similar. See (Conover, 1971) for more details.

ideological perspective, the fewer annotators we need to assess whether annotators make random guesses. By checking Figure 1 researchers can decide how many annotators are needed at the desired confidence level.

2.2. Reliability

We assess the reliability of the Vox Populi Annotation method by assessing whether the Vox Populi Intensity from one group of annotators is similar to the intensity from another group of annotators of the same size. The Vox Populi Annotation method is not reliable if intensity's magnitude changes greatly from one group of annotators to another group. Suppose 75% of annotators in a group label a sentence as Perspective A. The Vox Populi Annotation method is reliable if the same sentence is given to another group of annotators, and the number of annotators label the same sentence as Perspective A is still close to 75%.

However, the above assessment method may be fooled by *random* guessing. Consider the following two random guessing cases. In the first case, annotators make completely random guesses between two contrasting perspectives, either because they are under-trained or ideological perspectives are too hard to identify at the sentence level. Either way, two groups of such random-guessing annotators will consistently output Vox Populi Intensity 0.5 for every sentence. The magnitude of intensities is similar, but it does not mean the annotations are not random.

In the second case annotators keep making a biased decision, possibly because they are aware of the disproportion of two ideological perspectives in a corpus. Suppose two groups of annotators label a sentence to be the Israeli perspective 99% of the time. These two groups will label every sentence similarly, i.e., Vox Populi Intensity 0.99. However, we should not consider these Vox Populi Intensities as reliable because of superficial similarity resulting from biased guessing.

We choose Pearson's correlation coefficients to assess the reliability of the Vox Populi Annotation method. Given two sets of Vox Populi Intensities, $\{x_i\}$ and $\{y_i\}$, the Pearson correlation coefficient r is defined as follows,

$$r = \frac{n \sum_i x_i \sum_i y_i}{\sqrt{n \sum_i x_i^2 - (\sum_i x_i)^2} \sqrt{n \sum_i y_i^2 - (\sum_i y_i)^2}},$$

where n is the number of annotators. Correlation coefficients are positive and large when Vox Populi Intensity is positively correlated across different groups of annotators, and close to zero when Vox Populi Intensity is not related between different annotator groups. Correlation coefficients for the above two random cases will be zero because two groups of annotators make *independent* judgments (Casella and Berger, 2001). Therefore, the Vox Populi Annotation method is reliable if the correlation coefficients between two annotator groups are positive, high, and above zero.

3. Measuring Intensity of Ideological Perspectives

3.1. Annotation corpus and procedure

We randomly chose 250 sentences from the bitterlemons corpus (Lin et al., 2006), which consists of articles published on the website `http://bitterlemons.org/`. The website is set up to "contribute to mutual understanding [between Palestinians and Israelis] through the open exchange of ideas[2]." Every week an issue about the Israeli-Palestinian conflict is selected for discussion (e.g., "Disengagement: unilateral or coordinated?"), and a Palestinian editor and an Israeli editor each contribute one article addressing the issue. In addition, the Israeli and Palestinian editors invite one Israeli and one Palestinian to express their views on the issue (sometimes in the form of an interview), resulting in a total of four articles in a weekly edition.

We recruited annotators from Carnegie Mellon University students and staff. Participants were asked to label a sentence in the recruitment advertisement. Annotators signed a consent form that has been approved by the Institutional Review Board. A web-based interface displayed one sentence at a time, including the discussion topic and publishing date. Annotators were instructed to judge the sentence by making a forced binary choice on the question "Do you think the sentence is written from the Israeli or Palestinian perspective?". We encouraged participants to guess even when they were not sure. Eighteen of 26 participants finished the annotation study. Most of participants took one hour to finish the annotation study.

3.2. Annotation Results

The histogram of the Vox Populi Intensities of 250 sentences based on 18 annotators is shown in Figure 2. The intensity's distribution is bimodal, with one peak around 0.35 (i.e., more Palestinian) and one around 0.65 (i.e., more Israeli). The bimodal distribution suggests that two ideological perspectives in the bitterlemons corpus seem to be identifiable at the sentence level. If ideological perspectives could not be identified at the sentence level, annotators would mostly make random guesses, resulting in a distribution of Vox Populi Intensities closely centered around 0.5. The stretched distribution also suggests that our annotations contain sentences of varying intensities, which results in a much more rich language resource than simply strongly one-sided (i.e., all close to 0 or 1) or weak (i.e., all close to 0.5). Based on our analysis in Section 2.1., intensities greater than 0.75 and smaller than 0.25 are significantly not random in our annotations based on 18 annotators. Table 1 shows example sentences and their intensities of ideological perspectives.

3.3. Reliability Assessment

We calculated the Pearson correlation coefficients[3] of Vox Populi Intensities between two annotator groups to assess reliability. As described in Section 2.2., given $2n$ annotators, we randomly divided them into two groups of n annotators. For example, if we have 12 annotators, we randomly divide them into two 6-people annotator groups. We then calculated the Vox Populi Intensities of 250 sentences from each group, and computed the correlation coefficients between these two sets of 250 Vox Populi Intensities. We

[2] `http://www.bitterlemons.org/about/about.html`

[3] The results using rank-based correlation methods (Kendall's tau and Spearman's rho) are similar and thus omitted.

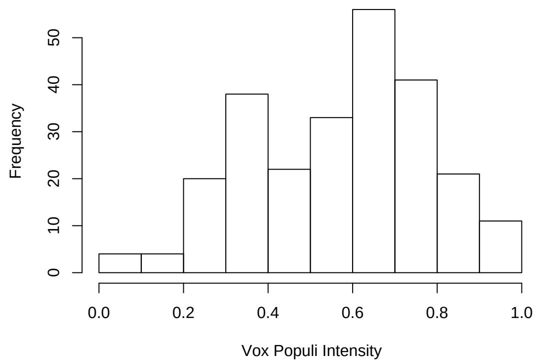

Figure 2: A histogram of Vox Populi Intensities of 250 sentences on the Israeli-Palestinian conflict. The larger the value, the more annotators judge a sentence to be written from the Israeli perspective.

repeatedly sample different $2n$ annotators, and reported the average of the 100 correlation coefficients.

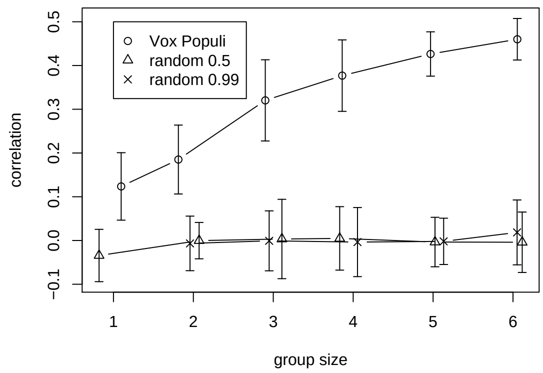

Figure 3: The correlation coefficients of Vox Populi Intensity and two random baselines as group sizes vary from one to six. We jittered the coordinate of group size to avoid the overlap between two random baselines.

We plotted the correlation coefficients between two group and two random guessing baselines in Figure 3. The correlation coefficients of the Vox Populi Annotation method differed significantly from 0 and two random baselines when the group size is large. The results suggested that Vox Populi Annotations, at least on the bitterlemons corpus, were unlikely to be randomly annotated and appears to be reliable. The positive correlation coefficients suggested that similar intensity estimates were likely to obtained no matter where annotator groups came from. The median and maximal pair-wise kappa statistic among 18 annotators on the 250 sentences were 0.10 and 0.44, respectively, which was very unsatisfactory according to (Carletta, 1996). As group sizes get larger, the aggregated Vox Populi Intensity becomes positively and highly correlated between two annotator groups even within each group two annotators may still still disagree with each other (i.e., low pair-wise kappa).

Vox Populi Intensity	Example
0.0769	The first is that Bush has placed the Palestinian-Israeli conflict squarely within the war against terrorism.
0.2307	This government, on the contrary, is trying in many ways, including through the so-called disengagement plan, to consolidate the occupation.
0.5384	This was not inevitable: as an election year approaches and the US sinks deeper into the Iraqi morass, Washington is simply not prepared to give high enough priority to the Israeli-Palestinian issue.
0.7692	Nor are security for Israelis and an end to terrorism–major topics of emphasis in Bush's presentation–likely to be achieved in this way.
0.9231	Palestinians in these ongoing debates have been basing their objections to the plan specifically on the argument that it contradicts the road map, for example on the issue of settlements.

Table 1: Five sentences and their Vox Populi Intensities of ideological perspectives. The larger the value, the more annotators judge a sentence to be written from the Israeli perspective.

4. Related Work

The Vox Populi Annotation method is not restricted to annotating to what degree a sentence conveys one of two contrasting perspectives. The Vox Populi Annotation method is a general methodology and may be applicable to other annotation tasks. As more computational linguistics are interested in more complex linguistic phenomena (e.g., intensity of subjectivity (Wiebe et al., 2005), political and social controversies (Meyers et al., 2007)), the Vox method Populi Annotation method can be a viable alternative for researchers to quantitatively measure these complex phenomena.

However, the Vox Populi Annotation method is not applicable to annotation tasks that require extensive linguistic knowledge or have little ambiguities:

- Annotation tasks that require extensive linguistic knowledge include, for example, predicate argument stricture in the Penn treebank (Meyers et al., 2007) and Rhetorical Structure Theory (RST) (Carlson et al., 2001). Because these annotation tasks require intensive training and constant monitoring, the cost of recruiting a large number of annotators becomes prohibitive. Besides, qualified candidates are very unlikely to be recruited from general public.

- Annotation tasks that have little ambiguities include, for example, named entities (Chinchor, 1997) and automatic speech recognition transcriptions (Fisher et al., 1986). Multiple annotators make little sense because they will label very similarly.

Our annotation study is about labeling intensity of bipolar ideological perspectives, but how about some annotation tasks that require more than two choices? For example, an annotation study may investigate the ideologies of different ethnic groups on immigration issues, and ask a group of annotators to decide if a sentence is written from Asian, Hispanic, or African ethnic groups' viewpoints (i.e., three categories). We can extend our reliability assessment in Section 2. to more than two choices. The exact binomial test for determining the number of annotators in Section 2.1. will be replaced by a multinomial test (Read and Cressie, 1988). The null hypothesis will not be a simple $\mu = 0.5$, and will be a multinomial vector that assumes every category is equally likely. The correlation coefficient for assessing the reliability of the Vox Populi Annotation method will be replaced by multivariate correlation (DuBois, 1957).

There has been annotation studies on measuring intensity, for example, the intensity of opinioned expressions (Wiebe et al., 2005). The annotation schemes in previous work mostly use the Likert Scale (Likert, 1932) and quantize intensity into discrete categories (e.g., low, medium, strong, and extreme). To have two annotators agree on each scale requires extensive training. Moreover, it is not trivial at all to transform annotations in Likert Scales to numerical values (Wu, 2007). On the contrary, Vox Populi Intensity is already a number and requires no transformation, which is important for evaluating computer programs that can output confidence scores.

The mathematical relationship between annotation group sizes and a sentence's intensity in Section 2.1. seems to be empirically observed. In a subjectivity annotation study (Wiebe et al., 2005),

> ... the difference between no subjectivity and a low-intensity private state might be highly debatable, but the difference between no subjectivity and a medium or high-intensity private sate is often much clearer.

The p-value formula based on the exact binomial test matches well the empirical observation. High intensity sentences are easier (i.e., requires fewer annotators) to be distinguished from random guessing than low intensity sentences.

There has been work using a large number of annotators to reduce annotators' bias (Eugenio and Glass, 2004), that is, individual annotators may have different preferences to label one category more than the other category. Incidentally, the same number of 18 annotators as ours were recruited in the study (Artstein and Poesio, 2005a). We explicitly determined the number of annotators based on the analysis in Section 2.1., and not simply chose a big number.

One seeming obstacle to the Vox Populi Annotation method is a large number of annotators. How can we afford so many annotators? While most annotation studies in computational linguistics recruit few annotators, many "annotation" tasks in other fields have begun to "recruit" a huge number of people, i.e., Crowd-sourcing (Hoew, 2006). Millions of Internet users have constantly labeled web pages (e.g., Delicious[4]), photos (e.g., Flicker[5]), and videos (e.g., YouTube[6]) without being paid. ESP game (von Ahn and Dabbish, 2004) and Google Image Labeler[7] use games to quickly collect high quality image annotations. With right kinds of incentive mechanism and annotation platforms, annotation studies in computational linguistics are likely to replicable these success stories in other fields. The Vox Populi Annotation method is not for every annotation task, but for those annotation tasks that require little training, this paper offer guidelines on selecting number of annotators and assessing reliability. Recently there has been an annotation study on sentiment conducted on Amazon Mechanical Turk (Barr and Cabrera, 2006), a commercial web service that facilitates large number of annotators.

5. Conclusion

We annotated the intensity of ideological perspectives expressed in 250 sentences extracted from articles on the Israeli-Palestinian conflict. We estimated the intensity of bipolar perspectives by aggregating binary judgments from multiple annotators. We measured intensity repeatedly from different groups of annotators, and found that the magnitude of intensity was highly correlated, which suggested that the Vox Populi Annotation method was reliable across different annotator groups, at least for the task of measuring the intensity of ideological perspectives.

Acknowledgements

We would like to thank the anonymous reviewers for their valuable suggestions for improving this paper, and Janyce Wiebe for helpful discussions. This research was supported in part by the National Science Foundation (NSF) under Grant No. IIS-0205219.

6. References

Ron Artstein and Massimo Poesio. 2005a. Bias decreases in proportion to the number of annotators. In *Proceedings of the Workshop on Formal Grammar – the Mathematics of Language*, pages 141–150.

Ron Artstein and Massimo Poesio. 2005b. Kappa^3 = Alpha (or Beta). Technical Report CSM-437, Department of Computer Science, University of Essex, September.

Jeff Barr and Luis Felipe Cabrera. 2006. AI gets a brain. *Queue*, 4(4):24–29, May.

[4] `http://del.icio.us/`

[5] `http://www.flickr.com/`

[6] `http://www.youtube.com/`

[7] `http://images.google.com/imagelabeler/`

Jean Carletta. 1996. Assessing agreement on classification tasks: The kappa statistics. *Computational Linguistics*, 22(2).

Lynn Carlson, Daniel Marcu, and mary Ellen Okurowski. 2001. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *Proceedings of the Second SIGDial Workshop on Discourse and Dialogue*, pages 1–10.

George Casella and Roger L. Berger. 2001. *Statistical Inference*. Duxbury Press, second edition.

Nancy Chinchor. 1997. MUC-7 named entity task definition. In *Proceedings of the Seventh Message Understanding Conference*.

William J. Conover. 1971. *Practical Nonparametric Statistics*. John Wiley & Sons.

Philip H. DuBois. 1957. *Multivariate correlational analysis*. Harper.

Barbara Di Eugenio and Michael Glass. 2004. The kappa statistics: a second look. *Computational Linguistics*, 30(1):95–101, March.

W.M. Fisher, G.R. Doddington, and K.M. Goudie-Marshall. 1986. The DARPA speech recognition research database: specifications and status. In *Proceedings of the DARPA Speech Recognition Workshop*.

Jeff Hoew. 2006. The rise of crowdsourcing. *Wired Magazine*, 14.

Rensis Likert. 1932. A technique for the measurement of attitudes. Archives of Psychology 140, Columbia University, New York City.

Wei-Hao Lin, Theresa Wilson, Janyce Wiebe, and Alexander Hauptmann. 2006. Which side are you on? identifying perspectives at the document and sentence levels. In *Proceedings of Tenth Conference on Natural Language Learning (CoNLL)*.

Adam Meyers, Nancy Ide, Ludovic Denoyer, and Yusuke Shinyama. 2007. The shared corpora working group report. In *Proceedings of the Linguistic Annotation Workshop*.

Rimothy R.C. Read and Noel A.C. Cressie. 1988. *Goodness-of-Fit Statistics for Discrete Multivariate Data*. Springer-Verlag.

Luis von Ahn and Laura Dabbish. 2004. Labeling images with a computer game. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 319–326.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2–3):165–210.

Chien-Ho Wu. 2007. An empirical study on the transformation of likert-scale data to numerical scores. *Applied Mathematical Sciences*, 1(58):2851–2862.