An Annotation Scheme for Automated Bias Detection in Wikipedia

Livnat Herzig, Alex Nunes and Batia Snir
Computer Science Department
Brandeis University
Waltham, MA, U.S.A.
lherzig, nunesa, bsnir @brandeis.edu

Abstract

BiasML is a novel annotation scheme with the purpose of identifying the presence as well as nuances of biased language within the subset of Wikipedia articles dedicated to service providers. Whereas Wikipedia currently uses only manual flagging to detect possible bias, our scheme provides a foundation for the automating of bias flagging by improving upon the methodology of annotation schemes in classic sentiment analysis. We also address challenges unique to the task of identifying biased writing within the specific context of Wikipedia's neutrality policy. We perform a detailed analysis of inter-annotator agreement, which shows that although the agreement scores for intra-sentential tags were relatively low, the agreement scores on the sentence and entry levels were encouraging (74.8% and 66.7%, respectively). Based on an analysis of our first implementation of our scheme, we suggest possible improvements to our guidelines, in hope that further rounds of annotation after incorporating them could provide appropriate data for use within a machine learning framework for automated detection of bias within Wikipedia.

1 Introduction

BiasML is an annotation scheme directed at detecting bias in the Wikipedia pages of service providers. Articles are judged as biased or non-biased at the sentential and document levels, and annotated on the intra-sentential level for a number of lexical and structural features.

2 Motivation and Background

2.1 Motivation

Neutral Point of View (NPOV) is one of three core tenets of Wikipedia's content policy. Wikipedia describes NPOV as "representing fairly, proportionately, and as far as possible without bias, all significant views that have been published by reliable sources" (Wikipedia, 2011a).

The collaborative design of Wikipedia is such that anyone can submit content, and so the detection and flagging of bias within articles is an essential and ongoing task in maintaining the quality and utility of Wikipedia. Currently, NPOV is enforced manually via the same open process that creates content on the site. Users can flag pages with suspect content as containing a "NPOV dispute". This is problematic: definitions of bias vary from editor to editor, and accusations of bias can themselves come from a biased perspective. Additionally, this practice is weighted towards the attention of Wikipedia users, such that the scrutiny an article receives is proportional to its broader popularity. For example, though the pages for Land of Israel and restaurant franchise Fresh to Order have both been flagged for NPOV disputes, they have been edited 1,480 and 46 times by 536 and 22 users, respectively (Wikipedia, 2011b; Wikipedia, 2011c). The average Wikipedia page receives just under 20 edits (Wikipedia, 2011d).

In light of this, an automated pass at bias detection is highly desirable. Instead of wholesale reliance on human editors, a system based on our annotation scheme could serve as an initial

Proceedings of the Fifth Law Workshop (LAW V), pages 47–55,
Portland, Oregon, 23-24 June 2011.

filter in monitoring user contributions. If integrated into the Wikipedia framework, this system could aid in the regulation of NPOV policy violations, e.g. tracking repeat offenders. With this goal in mind we have designed BiasML to flag NPOV issues in a specific subset of Wikipedia articles. We have constrained our task to the pages of service providers such as small businesses, schools, and hospitals. As a genre, the pages of service providers are especially worthy of scrutiny because they are both less likely to be closely vetted, and more likely to be edited by someone with a commercial interest in the reputation of the organization.

In addition, service provider pages are particularly appropriate for automatic POV-flagging because the bias complaints leveled against them tend to be much more systematic and objective compared with those of an especially controversial or divisive topic.

2.2 Background

Sentiment analysis efforts usually rely on the prior polarity of words (their polarity out of context). For example, Turney (2002) proposes a method to classify reviews as "recommended"/"not recommended", based on the average semantic orientation of the review. Semantic orientation is the mutual information measure of selected phrases with the word *excellent* minus their mutual information with the word *poor*. However, as Wilson et al. (2005) point out, even using a lexicon of positive/negative words marked for their prior polarity is merely a starting point, since a word's polarity in context might differ from its prior polarity.

The distinction between prior and contextual polarity is crucial for detecting bias, since words with a prior positive/negative polarity may or may not convey bias, depending on their context. Notably, the inverse is also true - generally neutral words can be used to create a favorable tone towards a sentence's topic, thereby expressing bias. An example of the latter case are the words *own* and *even* in the sentence *The hospital has its own pharmacy, maternity ward, and even a morgue*. Though generally neutral, their usage here contributes to the sentence's overall non-neutrality. In order to deal with contextual polarity, Wilson et al. propose a two-stage process that first uses clues marked with contextual polarity to determine whether the phrases containing these clues are polar or neutral. The second stage then determines the actual polarity of the phrases deemed non-neutral.

However, Wilson et al.'s approach would not suit our task of bias detection in Wikipedia, as the abovementioned example, taken from a Wikipedia entry, shows. Blatant expression of opinions or emotions is rare in the Wikipedia entries of service providers. Words which explicitly convey that an opinion/emotion is being expressed are rarely used (e.g. *I think*). Rather, bias is introduced either in more subtle ways (e.g. using words that are usually neutral) or in ways that differ from the ones addressed by previous approaches. For example, bias is introduced by preceding positive information about the provided service by phrases such as *it is widely believed*. Clearly, this phrase does not have contextual polarity, but it does introduce bias.

Within the realm of Wikipedia, phrases that create an impression that something specific and meaningful has been said when only a vague or ambiguous claim has been communicated, such as *it is widely believed*, are referred to as *weasels* (Wikipedia, 2011e). The recent CoNLL-2010 shared task (Farkas et al., 2010), aimed at detecting uncertainty cues in texts, focused on these phrases in trying to determine whether sentences contain uncertain information. In the same vein, we include weasel words as part of our annotation scheme to detect bias.

Finally, as Blitzer et al. (2007) point out, although the typical word-level analysis captures the finer-grained aspects of sentiment language, it falls short in capturing broader structurally or contextually-based bias. Bias can also be introduced by repetitive usage of words that in typical usage do not have prior polarity, but when used in a repetitive manner, create a favorable depiction of a sentence's topic. This cannot be captured by approaches such as those of Wilson et al. or Turney.

To tackle cases like those described above, our annotation scheme extends beyond lexical tags, and includes tags that capture dependencies between a word and its context, as well as tags that are aimed at capturing subtle expressions of bias.

3 Method

3.1 Corpus Selection and Preparation

The POV Wikipedia entries were selected from Wikipedia's list of entries that are classified as "NPOV dispute". Roughly 6,000 of the more than 3 million existing Wikipedia entries have been flagged this way (Wikipedia, 2011f). We went over these entries using a "get random article" feature, choosing ones that met our service provider criterion, i.e., they were either about a specific product or a service provider. The neutral entries were selected via a search through pages of products/service providers on Wikipedia that were evaluated by us as neutral. Our corpus ultimately consisted of 22 POV entries and 11 NPOV ones.

3.2 Annotation Scheme

Annotation Procedure and Tags: The annotation was performed using the MAE annotation tool (Stubbs, 2011), which is compliant with LAF guidelines (Ide and Romary, 2006). The annotation scheme uses standoff annotation and includes tagging on multiple levels - tagging biased words and linguistic structures; tagging the neutrality of each sentence; tagging the overall neutrality of the entry. The annotator is instructed to read through each sentence, and decide if it is written in a neutral point of view or not. At this point in the annotation process, a sentence is considered non-neutral if it is written in a non-neutral tone, or if it favors/disfavors its topic (regardless of whether the sentence is sourced). If a sentence is deemed neutral, it is tagged with a sentential level tag SENTENCE_POV, with the attribute NPOV, and no further tagging of it is required.

In the alternate case that a sentence is judged to contain non-neutral language, the annotator is asked to look for words/phrases that should be tagged with the word/phrase level tags (elaborated below) only within the scope of the current sentence. After tagging the word/phrase level tags, the sentence should be evaluated for its neutrality, and tagged SENTENCE_POV with one of two possible attributes (POV or NPOV), depending on the word/phrase level tags it has. After all the sentences are tagged with the SENTENCE_POV tag, the entire entry is tagged with the ENTRY_POV tag, whose attribute values are numeric, ranging between 1 and 4, where 1 is completely neutral and 4 is clearly non-neutral (i.e., written as an advertisement).

The annotation scheme is comprised of 4 word/phrase level extent tags that aim to capture biased language - POLAR_PHRASE, WEASEL, REPETITION, and PERSONAL_TONE. The POLAR_PHRASE tag is used to mark words/phrases that are used to express favor or disfavor within the sentential context, and contribute to the non-neutrality of the sentence. The annotator is advised to examine whether replacing the suspected word(s) results in a more neutral version of the sentence, without losing any of the sentence's content. If so, the word(s) should be tagged as POLAR_PHRASE (with a positive or negative attribute). For example, in the sentence *The new hospital <u>even</u> has a morgue*, *even* is tagged with the POLAR_PHRASE tag (the attribute value is positive), and the entire sentence's SENTENCE_POV tag receives the attribute POV.

The PERSONAL_TONE tag is used to tag words/phrases that convey a personal tone, which is commonly used in advertisements but is inappropriate in encyclopedic entries. The possible attribute values are first person (e.g. *we, our*), second person (e.g. *you, your*) and other (e.g. *here*). The REPETITION tag is used for two possible cases - when similar words are unnecessarily used to describe the same thing, all words except the first one should be considered a repetition; when there is unnecessary repetition that does not add new information (i.e., it is not elaboration, but mere repetition) about the service the service provider offers, or praise of the service provider, the repeated elements

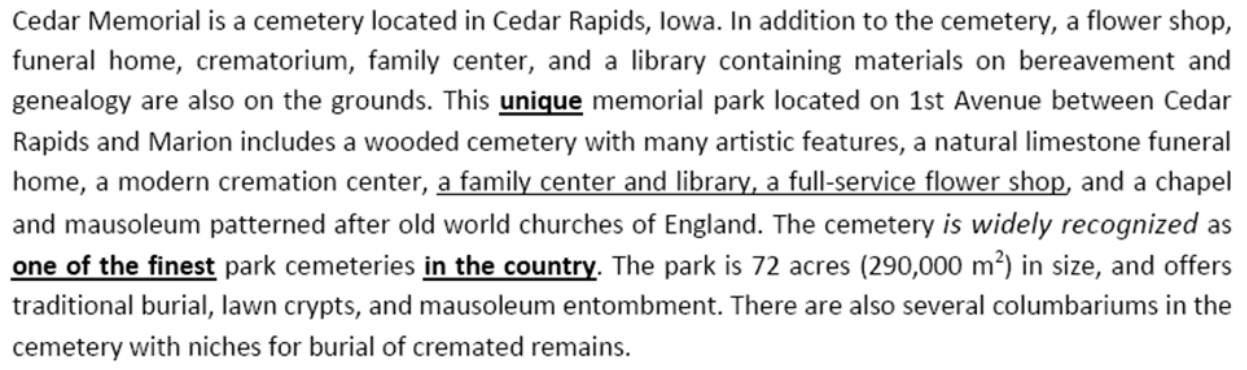

Figure 1: An annotated Wikipedia entry - POLAR_PHRASEs are underlined in bold, all of the positive type; WEASEL is italicized, and is of the pro type; REPETITION is underlined, receiving the attribute value 3. SENTENCE_POV for sentences no. 1, 2, 5 & 6 is NPOV, while it is POV for sentences no. 3 & 4. The ENTRY_POV is 3, which corresponds to POV.

should be considered repetition. For both cases, the attribute value will be the numeric value representing the number of repeated elements. To illustrate the former type of REPETITION and the PERSONAL_TONE tag, consider the sentence *The councils work to enhance and improve the quality of your local health service. Improve* is a case of REPETITION, since there is no need for both *enhance* and *improve* (the attribute value is 1). In addition, *your* is tagged with the PERSONAL_TONE tag (second person), and the sentence's SENTENCE_POV tag receives the attribute POV. The other type of REPETITION applies to cases where a sentence such as *The funeral home also offers a flower shop, crematorium, family center and library*, is subsequently followed by a sentence such as *This unique funeral home is built of natural limestone, and has a modern cremation center, a family center and library, a flower shop and a chapel*. While *unique* is tagged as a POLAR_PHRASE, the other underlined elements are all REPETITION, with the attribute value set to 3, since 3 elements are repeated unnecessarily, without adding new information. Note that although *crematorium* and *cremation center* refer to the same entity, it is not treated as a repetition, because the second mention adds that it is a modern crematorium. The second sentence's neutrality is therefore POV, while the first one's is NPOV.

As elaborated in the background section, weasel words also introduce bias, by presenting the appearance of support for statements while denying the reader the possibility to assess the viewpoint's source. These are usually general claims about what people think or feel, or what has been shown. These words/phrases are captured by the WEASEL tag. This tag has two possible attributes, pro, which captures "classic" WEASELs such as *is often credited*, and con, which would capture negative portrayal, as in *is never believed*. In contrast to the previously described word/phrase level tags, we also included a fifth tag, FACTIVE_PHRASE, which is inherently different. It is used to mark phrases that give objectivity to what is otherwise a biased description, usually a source. These phrases debias polar phrases and weasels.

The relation between a FACTIVE_PHRASE and the POLAR_PHRASE or WEASEL that it de-biases is captured by the LEGITIMIZE link tag. A sentence that was initially judged as non-neutral can eventually be tagged as NPOV, if each instance of its biased language is backed up by sources. Otherwise, it should be tagged as POV. For example, in the sentence *It is widely believed that John Smith started the tradition of pro-bono work.[1]*, the phrase *is widely believed* is tagged WEASEL, whereas *[1]* is tagged FACTIVE_PHRASE. In addition, a LEGITIMIZE tag will link these two elements,

resulting in an overall neutral sentence, since its biased language is backed up by a source. The SENTENCE_POV tag will therefore have the attribute value NPOV (whereas it would be POV if there were no FACTIVE_PHRASE). To further illustrate this point, consider the sentence *Jones and Sons ranked number one in The American Lawyer's Annual Survey. Number one* is tagged as a POLAR_PHRASE (positive), *The American Lawyer's Annual Survey* is a FACTIVE_PHRASE, and there is a LEGITIMIZE link between them. The entire SENTENCE_POV tag's neutrality is therefore NPOV. This is in contrast to the sentence *Jones and Sons are the number one law firm* in Boston., which would have the attribute value POV, because its polar phrases have no factive phrase to back them up. Our framework also enables tagging a sentence as POV even if none of the possible tags apply to them. See Figure 1 for an example of an annotated entry.

BiasML Innovations: The annotation scheme elaborated above is an innovative yet practical answer to the theoretical linguistic considerations of sentiment analysis within the genre of Wikipedia. As previously mentioned, our scheme improves upon approaches that rely upon prior polarity (e.g. Turney, 2002) by identifying cases of biased language that stem from intra-sentential and cross-sentential dependencies, rather than isolated words. Our POLAR_PHRASE tag resembles phrases with non-neutral contextual polarity that Wilson et al.'s (2005) approach introduces, but it captures cases that their approach does not - namely, generally neutral words that nevertheless make a sentence biased.

Another innovation of our framework is enabling the legitimization of weasel words. Whereas the CoNLL-2010 shared task (Farkas et al., 2010) annotated all occurrences of weasels as uncertainty markers, we acknowledge the possibility of sources (e.g. citations) that actually nullify the weasel.

The multiple-level discourse association of our tag scheme also allows observation of shifts in polarity within the larger discourse of the article. The sentence-level POV tag allows the annotator to identify the overall neutrality of each sentence, thus producing a landscape of how biased language is distributed across the article. This landscape not only provides an indicator of where to look for contextual clues and dependencies among more local tags, but it is particularly relevant to Wikipedia's wiki platform, where it is likely that different authors contributed to different portions of the article, making it more prone to variance in biased tone.

While developing this scheme, we wanted to make sure it tapped into the capacity of the annotator to identify both subjective language use and objective linguistic phenomena. While tags like PERSONAL_TONE and WEASEL require the annotator to mark precise occurrences of language, the sentence and document-level POV tags allow the annotator to identify point of view without having to explicitly point to a specific linguistic structure. To preserve the value of the human annotator's subjective judgments, our scheme permitted the co-occurrence of a sentence or document POV tag with the absence of any local lexical tags. This allowed our scheme to recognize the difficult cases in sentiment analysis where one intuitively senses opinionated language, but is unable to formally define what makes it so.

Another aim of our work was to develop a scheme that captured the way information is portrayed in Wikipedia, while avoiding judgment on what information is actually communicated. A significant source of dispute within Wikipedia is disagreement as to the veracity of an article's content; however, identification of this is truly a different task then the one we have defined here. In order to tease apart these distinct types of evaluation, annotators were instructed to identify citations that legitimize statements that are potentially POV, but not to consider the truthfulness of the statement or validity of the source when tagging.

4 Results

Our corpus of 33 articles of varying degrees of neutrality was distributed among three annotators, each annotator receiving 2/3 of the entire

corpus. The articles were presented as plain text in the annotation environment, and were stripped of images, titles, section headings, or other information extraneous to the main body of the text (inline references, however, were preserved). The annotators were graduate linguistics students. Their training consisted of a brief information session on the motivation of our work, a set of annotation guidelines, and optional question and answer sessions. Adjudication of the annotation was performed with the MAI adjudication tool (Stubbs, 2011).

4.1 Tag Analysis

For each tag, an average percent agreement score was calculated (for extents and attributes) per document, then averaged to get the agreement over all documents in the corpus. Note that extent agreement was defined as strictly as possible, requiring an exact character index match, meaning cases of overlap would not be considered agreement (e.g. *best* and *the best* would not be a match, even if they referred to the same instance of *best*). The percent agreement scores are displayed in Table 1. Note that calculations were not performed for the LEGITMIZE link tag, because it relies on the extent of other tags.

Tag	% Extent Agreement	% Attribute Agreement
POLAR_PHRASE	6.5	60
FACTIVE_PHRASE	9.3	NA
WEASEL	4.9	13.6
REPETITION	0	0
PERSONAL_TONE	33	57.1
SENTENCE_POV	94.6	74.8
ENTRY_POV	97	66.7

Table 1: Tag Analysis of IAA: Mean % Agreement

Agreement is notably stronger among the higher level tags, ENTRY_POV and SENTENCE_POV. For the ENTRY_POV neutrality attribute, we had decided to measure overall Entry_POV neutrality along a 4-point scale, after noticing our own hesitation to assign the same tag to both slightly preferential and flagrantly biased entries. However, this more nuanced system was at odds with our original objective of creating an annotation scheme for use in a binary classification of bias. Though it might manifest to different degrees, bias either is or is not present within an entry. Our intention in collapsing the scale after the fact was to recover a more organic division in Entry_POV judgments. With the built-in 4-way division, inter-annotator agreement on Entry_POV attributes stood at 42.42%. This number rose considerably when the scale was reduced to a 2-way division. To reflect the notion that any bias is unacceptable, we chose to divide ENTRY_POV into two groups: not-at-all-biased (ENTRY_POV=1) and containing bias (ENTRY_POV>1). This division yielded an inter-annotator agreement of 66.7%. In the case of the SENTENCE_POV attribute, which is binary, agreement on neutrality is even higher at 74.8%.

The strength of scores for attributes at the sentence and document levels suggest that annotators had similar perceptions of what kinds of discourse entailed a bias not fit for an encyclopedic entry. This in turn suggests that there is conceptual validity in our task on a higher level, as well as validity in how that concept was defined and conveyed to annotators.

Interestingly, agreement numbers decline for the intra-sentential tags. Both POLAR_PHRASE and PERSONAL_TONE have attribute agreement scores at or near 60%, but PERSONAL_TONE has an extent agreement of 33%, while POLAR_PHRASE has only 6.5% for extent. WEASEL and REPETITION have low scores for both extent and attribute, with REPETITION being 0% for both (note that extent agreement is a prerequisite for attribute agreement). FACTIVE_PHRASE also has low extent agreement, making extent agreement generally low across the board for intra-sentential tags.

Attribute agreement is expected to be high for the intra-sentential tags, given that attributes are almost always positive (pro/positive) within the service provider genre. Based on the adjudication process, we suspect that the main contributor to instances of attribute disagreement for these tags was simply a failure

on the annotators' part to specify the attribute at all, perhaps because they encountered mainly positive/pro instances of POLAR_PHRASEs/WEASELs, thereby forgetting that an attribute is relevant. The annotators also reported confusion about cases where a generally negative word/phrase is used to support or promote the article's topic (in these cases, the attribute should be positive).

For POLAR_PHRASE, the lack of extent agreement is not entirely unexpected, as this tag was difficult to define. As previously discussed, we chose not to use a lexicon of positive/negative words with their prior polarity, because a word's polarity in these documents was highly contingent upon its context and particular usage. During adjudication, it was observed that one of the annotators consistently marked any term that was generally positive as a POLAR_PHRASE. For example, the word *modern* was chosen when used to describe *architecture*. Although this word has some sort of positive connotation, it does not meet the substitution criteria outlined for POLAR_PHRASE in the guidelines (for a word to qualify as a POLAR_PHRASE, there should be a comparable substitution possible that would reduce the non-neutrality of the sentence without losing any of its content). This annotator had set his/her acceptability threshold for this tag too low, which resulted in over-selection. This could hopefully be avoided in future annotation efforts by more exposure to correct and incorrect examples of polar phrases.

Low extent agreement for the WEASEL and REPETITION tags appears to be a result of a poor understanding of what the tags are meant to capture. In the case of the WEASEL tag, annotators tended to mark anything that had an obscured source, such as, *being overlooked for the position* and *a number of executives*. Although the passive voice in the first example and the vague specification in the second one do obscure a source, they do not present support for the topic at hand, which is part of the WEASEL definition. To aid future annotation, it appears that further emphasis is needed to convey the fact that a WEASEL consists of a targeted word/phrase (and not just a lack of citation) that is used to conceal the source of a favorable or unfavorable statement. A lexicon would be useful in this case, as most weasels are covered by just a handful of common phrases or constructions. For example, *the famous* __ is a common WEASEL that was missed by all annotators throughout the corpus.

The poor performance for the REPETITION tag is probably a result of it not being just literal echo, but rather a recurrence of information used for promotional purposes. Like POLAR_PHRASE, this makes its definition rather subjective, and thus prone to different interpretations. Throughout the corpus, all annotators tended to miss the REPETITION we had identified in the gold standard, and there were also cases of annotators marking literal repetitions that did not match the guidelines' criteria. Although the linguistic phenomenon that the REPETITION tag was intended to capture is indeed indicative of bias (especially for service provider articles), it is relatively rare. Its rarity and elusiveness, combined with the fact that agreement was 0%, would motivate us to exclude this as a tag in future versions of the annotation scheme.

4.2 Annotator Analysis

Table 2 reports how each annotator compares to the gold standard (which was determined by the authors). Overall, annotator B clearly outperformed the other two, with both strong precision and recall scores. For all the intra-sentential tags with the exception of WEASEL, there seems to be a consistent trend where annotator B has the highest scores, a second annotator has somewhat lower scores (either A or C), and the third one has very low scores. This trend suggests that for each of these tags, a single annotator tended to pull down its agreement scores, though not consistently the same annotator. For example, annotator C performed relatively poorly on FACTIVE_PHRASE and PERSONAL_TONE, while the same was true for annotator A on the POLAR_PHRASE and REPETITION tags. For the higher level tags (SENTENCE_POV and ENTRY_POV), performance

was excellent for all annotators, which is consistent with the percent agreement scores from Table 1.

Tag	annotator_a pre., rec.	annotator_b pre., rec.	annotator_c pre., rec.
POLAR_PHRASE	0.2, 0.28	0.63, 0.89	0.55, 0.17
FACTIVE_PHRASE	0.29, 0.5	0.55, 0.86	0, 0
WEASEL	0.33, 0.28	0.85, 0.92	0.33, 0.6
REPETITION	0.06, 0.08	0.62, 1	0.44, 0.36
PERSONAL_TONE	0.64, 0.39	1, 1	0, 0
SENTENCE_POV	1, 0.97	1, 1	0.98, 0.97
ENTRY_POV	1, 1	1, 1	1, 1

Table 2: Per-Annotator Analysis: Precision and Recall

While the low individual scores on intra-sentential tags is disconcerting, the overall higher scores for annotator B are a positive indication that a decent understanding and execution of the scheme and guidelines are possible, and agreement could potentially improve greatly with better training for adherence to the guidelines in the case of the other two annotators.

4.3 Proposed Annotation Changes

Post-annotation analyses have provided a basis for changes to our annotation scheme, guidelines, and implementation process for the future. In addition to the changes to the guidelines we have suggested in the previous section, we believe that the greatest amount of improvement for our tag agreement could be achieved by conducting a training session for annotators, in which they study and then practice with positive and negative examples of the different tags. This would hopefully solidify understanding of the tagging scheme, since it became apparent during comparison with the gold standard that certain annotators had trouble with specific tags. Furthermore, it would be worth experimenting with less rigorous forms of extent matching, and perhaps allowing extents with a certain degree of overlap to qualify as agreement.

5 Conclusions and Future Work

The work presented here offers a new annotation scheme for the automatic detection of bias in the unique genre of Wikipedia entries. In addition to a tagset designed to identify linguistic characteristics associated with bias within an encyclopedic corpus, our scheme works beyond typical sentiment analysis approaches to capture cross-sentential linguistic phenomena that lead to encyclopedia bias. Strong agreement results for sentence and document levels bias tags (74.8% and 66.7%, respectively) indicate that there is conceptual validity in our task on a higher level, as well as validity in how that concept was defined and conveyed to annotators. While agreement for intra-sentential tags was lower, the fact that one annotator consistently scored high on agreement with the gold standard suggests that improved annotator training, and specification of unforeseen cases in the guidelines would provide more reliable annotator performance for these tags. It is our hope that upon implementing the suggested improvements outlined in this work, further rounds of annotation could provide appropriate data for use within a machine learning framework for automated detection of various sorts of bias within Wikipedia.

Acknowledgments

We would like to thank James Pustejovsky, Lotus Goldberg and Amber Stubbs for feedback on earlier versions of this paper and helpful advice along the execution of this project. We would also like to thank three anonymous reviewers for their comments.

References

John Blitzer, Mark Drezde , and Fernando Pereira. 2007. Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaptation for Sentiment Classification. *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, 187–205. Prague, Czech Republic.

Richard Farkas, Veronika Vincze, Gyorgy Mora, Janos Csirik and Gyorgy Szarvas. 2010. The CoNLL-2010 Shared Task: Learning to Detect Hedges and their Scope in Natural Language Text.

Proceedings of the Fourteenth Conference on Computational Natural Language Learning: Shared Task, 1–12. Uppsala, Sweden.

Nancy Ide and Laurent Romary. 2006. Representing Linguistic Corpora and Their Annotations. *Proceedings of the Fifth Language Resources and Evaluation Conference*, Genoa, Italy.

Amber Stubbs. 2011. MAE and MAI: Lightweight Annotation and Adjudication Tools. *Proceedings of the Fifth Linguistic Annotation Workshop. LAW V*. Portland, Oregon.

Peter D. Turney. 2002. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification fo Reviews. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 417–424. Philadelphia, Pennsylvania.

Wikipedia. 2011a. http://en.wikipedia.org/wiki/Wikipedia:Neutral_point_of_view. Accessed May 5, 2011.

Wikipedia. 2011b. http://toolserver.org/~soxred93/articleinfo/index.php?%20article=Land_of_Israel&lang=en&wiki=wikipedia. Accessed May 5, 2011.

Wikipedia. 2011c. http://toolserver.org/~soxred93/articleinfo/index.php?%20article=Fresh_to_Order&lang=en&wiki=wikipedia. Accessed May 5, 2011.

Wikipedia. 2011d. http://en.wikipedia.org/wiki/Special:Statistics. Accessed May 5, 2011.

Wikipedia. 2011e. http://en.wikipedia.org/wiki/Weasel_word. Accessed May 5, 2011.

Wikipedia. 2011f. http://en.wikipedia.org/wiki/Category:NPOV_disputes. Accessed May 5, 2011.

Theresa Wilson, Janyce Wiebe, and Paul Hoffman. 2005. Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. *Joint Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing*, 347–354. Vancouver, Canada.