Linguistic Models for Analyzing and Detecting Biased Language

Marta Recasens
Stanford University
recasens@google.com

Cristian Danescu-Niculescu-Mizil
Stanford University
Max Planck Institute SWS
cristiand@cs.stanford.edu

Dan Jurafsky
Stanford University
jurafsky@stanford.edu

Abstract

Unbiased language is a requirement for reference sources like encyclopedias and scientific texts. Bias is, nonetheless, ubiquitous, making it crucial to understand its nature and linguistic realization and hence detect bias automatically. To this end we analyze real instances of human edits designed to remove bias from Wikipedia articles. The analysis uncovers two classes of bias: *framing bias*, such as praising or perspective-specific words, which we link to the literature on subjectivity; and *epistemological bias*, related to whether propositions that are presupposed or entailed in the text are uncontroversially accepted as true. We identify common linguistic cues for these classes, including factive verbs, implicatives, hedges, and subjective intensifiers. These insights help us develop features for a model to solve a new prediction task of practical importance: given a biased sentence, identify the bias-inducing word. Our linguistically-informed model performs almost as well as humans tested on the same task.

1 Introduction

Writers and editors of reference works such as encyclopedias, textbooks, and scientific articles strive to keep their language **unbiased**. For example, Wikipedia advocates a policy called *neutral point of view (NPOV)*, according to which articles should represent "fairly, proportionately, and as far as possible without bias, all significant views that have been published by reliable sources" (Wikipedia, 2013b). Wikipedia's style guide asks editors to use nonjudgmental language, to indicate the relative prominence of opposing points of view, to avoid presenting uncontroversial facts as mere opinion, and, conversely, to avoid stating opinions or contested assertions as facts.

Understanding the linguistic realization of bias is important for linguistic theory; automatically detecting these biases is equally significant for computational linguistics. We propose to address both by using a powerful resource: edits in Wikipedia that are specifically designed to remove bias. Since Wikipedia maintains a complete revision history, the edits associated with NPOV tags allow us to compare the text in its biased (before) and unbiased (after) form, helping us better understand the linguistic realization of bias. Our work thus shares the intuition of prior NLP work applying Wikipedia's revision history (Nelken and Yamangil, 2008; Yatskar et al., 2010; Max and Wisniewski, 2010; Zanzotto and Pennacchiotti, 2010).

The analysis of Wikipedia's edits provides valuable linguistic insights into the nature of biased language. We find two major classes of bias-driven edits. The first, **framing bias**, is realized by subjective words or phrases linked with a particular point of view. In (1), the term *McMansion*, unlike *homes*, appeals to a negative attitude toward large and pretentious houses. The second class, **epistemological bias**, is related to linguistic features that subtly (often via presupposition) focus on the believability of a proposition. In (2), the assertive *stated* removes the bias introduced by *claimed*, which casts doubt on Kuypers' statement.

(1) a. Usually, smaller cottage-style houses have been demolished to make way for these **McMansions**.
b. Usually, smaller cottage-style houses have been demolished to make way for these **homes**.

(2) a. Kuypers **claimed** that the mainstream press in America tends to favor liberal viewpoints.
b. Kuypers **stated** that the mainstream press in America tends to favor liberal viewpoints.

Bias is linked to the lexical and grammatical cues identified by the literature on subjectivity (Wiebe et al., 2004; Lin et al., 2011), sentiment (Liu et al., 2005; Turney, 2002), and especially stance

or "arguing subjectivity" (Lin et al., 2006; Somasundaran and Wiebe, 2010; Yano et al., 2010; Park et al., 2011; Conrad et al., 2012). For example, like stance, framing bias is realized when the writer of a text takes a particular position on a controversial topic and uses its metaphors and vocabulary. But unlike the product reviews or debate articles that overtly use subjective language, editors in Wikipedia are actively trying to avoid bias, and hence biases may appear more subtly, in the form of covert framing language, or presuppositions and entailments that may not play as important a role in other genres. Our linguistic analysis identifies common classes of these subtle bias cues, including factive verbs, implicatives and other entailments, hedges, and subjective intensifiers.

Using these cues could help automatically detect and correct instances of bias, by first finding biased phrases, then identifying the word that introduces the bias, and finally rewording to eliminate the bias. In this paper we propose a solution for the second of these tasks, identifying the bias-inducing word in a biased phrase. Since, as we show below, this task is quite challenging for humans, our system has the potential to be very useful in improving the neutrality of reference works like Wikipedia. Tested on a subset of non-neutral sentences from Wikipedia, our model achieves 34% accuracy—and up to 59% if the top three guesses are considered—on this difficult task, outperforming four baselines and nearing humans tested on the same data.

2 Analyzing a Dataset of Biased Language

We begin with an empirical analysis based on Wikipedia's bias-driven edits. This section describes the data, and summarizes our linguistic analysis.[1]

2.1 The NPOV Corpus from Wikipedia

Given Wikipedia's strict enforcement of an NPOV policy, we decided to build the **NPOV corpus**, containing Wikipedia edits that are specifically designed to remove bias. Editors are encouraged to identify and rewrite biased passages to achieve a more neutral tone, and they can use several NPOV tags to mark biased content.[2] Articles tagged this way fall into Wikipedia's category of *NPOV disputes*.

Data	Articles	Revisions	Words	Edits	Sents
Train	5997	2238K	11G	13807	1843
Dev	653	210K	0.9G	1261	163
Test	814	260K	1G	1751	230
Total	7464	2708K	13G	16819	2235

Table 1: Statistics of the NPOV corpus, extracted from Wikipedia. (*Edits* refers to bias-driven edits, i.e., with an NPOV comment. *Sents* refers to sentences with one-word bias-driven edit.)

We constructed the NPOV corpus by retrieving all articles that were or had been in the NPOV-dispute category[3] together with their full revision history. We used Stanford's CoreNLP tools[4] to tokenize and split the text into sentences. Table 1 shows the statistics of this corpus, which we split into training (train), development (dev), and test. Following Wikipedia's terminology, we call each version of a Wikipedia article a *revision*, and so an article can be viewed as a set of (chronologically ordered) revisions.

2.2 Extracting Edits Meant to Remove Bias

Given all the revisions of a page, we extracted the changes between pairs of revisions with the word-mode *diff* function from the Diff Match and Patch library.[5] We refer to these changes between revisions as *edits*, e.g., *McMansion > large home*. An edit consists of two strings: the old string that is being replaced (i.e., the before form), and the new modified string (i.e., the after form).

Our assumption was that among the edits happening in NPOV disputes, we would have a high density of edits intended to remove bias, which we call *bias-driven edits*, like (1) and (2) from Section 1. But many other edits occur even in NPOV disputes, including edits to fix spelling or grammatical errors, simplify the language, make the meaning more precise, or even vandalism (Max

[1] The data and bias lexicon we developed are available at http://www.mpi-sws.org/~cristian/Biased_language.html

[2] {{POV}}, {{POV-check}}, {{POV-section}}, etc. Adding these tags displays a template such as "The neutrality of this article is disputed. Relevant discussion may be found on the talk page. Please do not remove this message until the dispute is resolved."

[3] http://en.wikipedia.org/wiki/Category:All_NPOV_disputes

[4] http://nlp.stanford.edu/software/corenlp.shtml

[5] http://code.google.com/p/google-diff-match-patch

and Wisniewski, 2010). Therefore, in order to extract a high-precision set of bias-driven edits, we took advantage of the comments that editors can associate with a revision—typically short and brief sentences describing the reason behind the revision. We considered as bias-driven edits those that appeared in a revision whose comment mentioned *(N)POV*, e.g., *Attempts at presenting some claims in more NPOV way*; or *merging in a passage from the researchers article after basic NPOVing*. We only kept edits whose before and after forms contained five or fewer words, and discarded those that only added a hyperlink or that involved a minimal change (character-based Levenshtein distance < 4). The final number of bias-driven edits for each of the data sets is shown in the "Edits" column of Table 1.

2.3 Linguistic Analysis

Style guides talk about biased language in a prescriptive manner, listing a few words that should be avoided because they are flattering, vague, or endorse a particular point of view (Wikipedia, 2013a). Our focus is on analyzing actual biased text and bias-driven edits extracted from Wikipedia.

As we suggested above, this analysis uncovered two major classes of bias: epistemological bias and framing bias. Table 2 shows the distribution (from a sample of 100 edits) of the different types and subtypes of bias presented in this section.

(A) Epistemological bias involves propositions that are either commonly agreed to be true or commonly agreed to be false and that are subtly presupposed, entailed, asserted or hedged in the text.

1. **Factive verbs** (Kiparsky and Kiparsky, 1970) presuppose the truth of their complement clause. In (3-a) and (4-a), *realize* and *reveal* presuppose the truth of "the oppression of black people..." and "the Meditation technique produces...", whereas (3-b) and (4-b) present the two propositions as somebody's stand or an experimental result.

   (3) a. **He realized** that the oppression of black people was more of a result of economic exploitation than anything innately racist.
   b. **His stand was** that the oppression of black people was more of a result of economic exploitation than anything innately racist.

   (4) a. The first research **revealed** that the Meditation technique produces a unique state fact.
   b. The first research **indicated** that the Meditation technique produces a unique state fact.

Bias	Subtype	%
A. Epistemological bias		43
	- Factive verbs	3
	- Entailments	25
	- Assertives	11
	- Hedges	4
B. Framing bias		57
	- Intensifiers	19
	- One-sided terms	38

Table 2: Proportion of the different bias types.

2. **Entailments** are directional relations that hold whenever the truth of one word or phrase follows from another, e.g., *murder* entails *kill* because there cannot be murdering without killing (5). However, *murder* entails killing in an unlawful, premeditated way. This class includes implicative verbs (Karttunen, 1971), which imply the truth or untruth of their complement, depending on the polarity of the main predicate. In (6-a), *coerced into accepting* entails accepting in an unwilling way.

   (5) a. After he **murdered** three policemen, the colony proclaimed Kelly a wanted outlaw.
   b. After he **killed** three policemen, the colony proclaimed Kelly a wanted outlaw.

   (6) a. A computer engineer who **was coerced into accepting** a plea bargain.
   b. A computer engineer who **accepted** a plea bargain.

3. **Assertive verbs** (Hooper, 1975) are those whose complement clauses assert a proposition. The truth of the proposition is not presupposed, but its level of certainty depends on the asserting verb. Whereas verbs of saying like *say* and *state* are usually neutral, *point out* and *claim* cast doubt on the certainty of the proposition.

   (7) a. The "no Boeing" theory is a controversial issue, even among conspiracists, many of whom have **pointed out** that it is disproved by ...
   b. The "no Boeing" theory is a controversial issue, even among conspiracists, many of whom have **said** that it is disproved by...

   (8) a. Cooper says that slavery was worse in South America and the US than Canada, but **clearly states** that it was a horrible and cruel practice.
   b. Cooper says that slavery was worse in South America and the US than Canada, but **points out** that it was a horrible and cruel practice.

4. **Hedges** are used to reduce one's commitment to the truth of a proposition, thus avoiding any bold predictions (9-b) or statements (10-a).[6]

(9) a. Eliminating the profit motive **will decrease the** rate of medical innovation.
b. Eliminating the profit motive **may have a lower** rate of medical innovation.

(10) a. The lower cost of living in more rural areas means a **possibly** higher standard of living.
b. The lower cost of living in more rural areas means a higher standard of living.

Epistemological bias is bidirectional, that is, bias can occur because doubt is cast on a proposition commonly assumed to be true, or because a presupposition or implication is made about a proposition commonly assumed to be false. For example, in (7) and (8) above, *point out* is replaced in the former case, but inserted in the second case. If the truth of the proposition is uncontroversially accepted by the community (i.e., reliable sources, etc.), then the use of a factive is unbiased. In contrast, if only a specific viewpoint agrees with its truth, then using a factive is biased.

(B) Framing bias is usually more explicit than epistemological bias because it occurs when subjective or one-sided words are used, revealing the author's stance in a particular debate (Entman, 2007).

1. **Subjective intensifiers** are adjectives or adverbs that add (subjective) force to the meaning of a phrase or proposition.

(11) a. Schnabel himself did the **fantastic** reproductions of Basquiat's work.
b. Schnabel himself did the **accurate** reproductions of Basquiat's work.

(12) a. Shwekey's albums are arranged by many **talented** arrangers.
b. Shwekey's albums are arranged by many **different** arrangers.

2. **One-sided terms** reflect only one of the sides of a contentious issue. They often belong to controversial subjects (e.g., religion, terrorism, etc.) where the same event can be seen from two or more opposing perspectives, like the Israeli-Palestinian conflict (Lin et al., 2006).

(13) a. Israeli forces **liberated** the eastern half of Jerusalem.
b. Israeli forces **captured** the eastern half of Jerusalem.

(14) a. Concerned Women for America's major areas of political activity have consisted of opposition to gay causes, **pro-life** law...
b. Concerned Women for America's major areas of political activity have consisted of opposition to gay causes, **anti-abortion** law...

(15) a. Colombian **terrorist** groups.
b. Colombian **paramilitary** groups.

Framing bias has been studied within the literature on stance recognition and arguing subjectivity. Because this literature has focused on identifying which side an article takes on a two-sided debate such as the Israeli-Palestinian conflict (Lin et al., 2006), most studies cast the problem as a two-way classification of documents or sentences into *for*/positive vs. *against*/negative (Anand et al., 2011; Conrad et al., 2012; Somasundaran and Wiebe, 2010), or into one of two opposing views (Yano et al., 2010; Park et al., 2011). The features used by these models include subjectivity and sentiment lexicons, counts of unigrams and bigrams, distributional similarity, discourse relationships, and so on.

The datasets used by these studies come from genres that overtly take a specific stance (e.g., debates, editorials, blog posts). In contrast, Wikipedia editors are asked not to advocate a particular point of view, but to provide a balanced account of the different available perspectives. For this reason, overtly biased opinion statements such as "I believe that..." are not common in Wikipedia. The features used by the subjectivity literature help us detect framing bias, but we also need features that capture epistemological bias expressed through presuppositions and entailments.

3 Automatically Identifying Biased Language

We now show how the bias cues identified in Section 2.3 can help solve a new task. Given a biased sentence (e.g., a sentence that a Wikipedia editor has tagged as violating the NPOV policy), our goal in this new task is to identify the word that introduces bias. This is part of a potential three-step process for detecting and correcting biased language: (1) finding biased phrases, (2) identifying the word that introduces the bias, (3) rewording to eliminate the bias. As we will see below, it can be

[6] See Choi et al. (2012) for an exploration of the interface between hedging and framing.

hard even for humans to track down the sources of bias, because biases in reference works are often subtle and implicit. An automatic bias detector that can highlight the bias-inducing word(s) and draw the editors' attention to words that need to be modified could thus be important for improving reference works like Wikipedia or even in news reporting.

We selected the subset of sentences that had a single NPOV edit involving one (original) word. (Although the before form consists of only one word, the after form can be either one or more words or the null string (i.e., deletion edits); we do not use the after string in this identification task). The number of sentences in the train, dev and test sets is shown in the last column of Table 1.

We trained a logistic regression model on a feature vector for every word that appears in the NPOV sentences from the training set, with the bias-inducing words as the positive class, and all the other words as the negative class. The features are described in the next section.

At test time, the model is given a set of sentences and, for each of them, it ranks the words according to their probability to be biased, and outputs the highest ranked word (TOP1 model), the two highest ranked words (TOP2 model), or the three highest ranked words (TOP3 model).

3.1 Features

The types of features used in the logistic regression model are listed in Table 3, together with their value space. The total number of features is 36,787. The ones targeting framing bias draw on previous work on sentiment and subjectivity detection (Wiebe et al., 2004; Liu et al., 2005). Features to capture epistemological bias are based on the bias cues identified in Section 2.3.

A major split separates the features that describe the word under analysis (e.g., lemma, POS, whether it is a hedge, etc.) from those that describe its surrounding context (e.g., the POS of the word to the left, whether there is a hedge in the context, etc.). We define *context* as a 5-gram window, i.e., two words to the left of the word under analysis, and two to the right. Taking context into account is important given that biases can be context-dependent, especially epistemological bias since it depends on the truth of a proposition. To define some of the features like POS and grammatical relation, we used the Stanford's CoreNLP tagger and dependency parser (de Marneffe et al., 2006).

Features 9–10 use the list of hedges from Hyland (2005), features 11–14 use the factives and assertives from Hooper (1975), features 15–16 use the implicatives from Karttunen (1971), features 19–20 use the entailments from Berant et al. (2012), features 21–25 employ the subjectivity lexicon from Riloff and Wiebe (2003), and features 26–29 use the sentiment lexicon—positive and negative words—from Liu et al. (2005). If the word (or a word in the context) is in the lexicon, then the feature is true, otherwise it is false.

We also included a "bias lexicon" (feature 31) that we built based on our NPOV corpus from Wikipedia. We used the training set to extract the lemmas of words that were the before form of at least two NPOV edits, and that occurred in at least two different articles. Of the 654 words included in this lexicon, 433 were unique to this lexicon (i.e., recorded in neither Riloff and Wiebe's (2003) subjectivity lexicon nor Liu et al.'s (2005) sentiment lexicon) and represented many one-sided or controversial terms, e.g., *abortion, same-sex, execute*.

Finally, we also included a "collaborative feature" that, based on the previous revisions of the edit's article, computes the ratio between the number of times that the word was NPOV-edited and its frequency of occurrence. This feature was designed to capture framing bias specific to an article or topic.

3.2 Baselines

Previous work on subjectivity and stance recognition has been evaluated on the task of classifying documents as opinionated vs. factual, *for* vs. *against*, positive vs. negative. Given that the task of identifying the bias-inducing word of a sentence is novel, there were no previous results to compare directly against. We ran the following five baselines.

1. **Random guessing**. Naively returns a random word from every sentence.
2. **Role baseline**. Selects the word with the syntactic role that has the highest probability to be biased, as computed on the training set. This is the parse tree root (probability $p = .126$ to be biased), followed by verbal arguments ($p = .085$), and the subject ($p = .084$).

ID	Feature	Value	Description
1*	Word	<string>	Word *w* under analysis.
2	Lemma	<string>	Lemma of *w*.
3*	POS	{NNP, JJ, ...}	POS of *w*.
4*	POS – 1	{NNP, JJ, ...}	POS of one word before *w*.
5	POS – 2	{NNP, JJ, ...}	POS of two words before *w*.
6*	POS + 1	{NNP, JJ, ...}	POS of one word after *w*.
7	POS + 2	{NNP, JJ, ...}	POS of two words after *w*.
8	Position in sentence	{start, mid, end}	Position of *w* in the sentence (split into three parts).
9	Hedge	{true, false}	*w* is in Hyland's (2005) list of hedges (e.g., *apparently*).
10*	Hedge in context	{true, false}	One/two words) around *w* is a hedge (Hyland, 2005).
11*	Factive verb	{true, false}	*w* is in Hooper's (1975) list of factives (e.g., *realize*).
12*	Factive verb in context	{true, false}	One/two word(s) around *w* is a factive (Hooper, 1975).
13*	Assertive verb	{true, false}	*w* is in Hooper's (1975) list of assertives (e.g., *claim*).
14*	Assertive verb in context	{true, false}	One/two word(s) around *w* is an assertive (Hooper, 1975).
15	Implicative verb	{true, false}	*w* is in Karttunen's (1971) list of implicatives (e.g., *manage*).
16*	Implicative verb in context	{true, false}	One/two word(s) around *w* is an implicative (Karttunen, 1971).
17*	Report verb	{true, false}	*w* is a report verb (e.g., *add*).
18	Report verb in context	{true, false}	One/two word(s) around *w* is a report verb.
19*	Entailment	{true, false}	*w* is in Berant et al.'s (2012) list of entailments (e.g., *kill*).
20*	Entailment in context	{true, false}	One/two word(s) around *w* is an entailment (Berant et al., 2012).
21*	Strong subjective	{true, false}	*w* is in Riloff and Wiebe's (2003) list of strong subjectives (e.g., *absolute*).
22	Strong subjective in context	{true, false}	One/two word(s) around *w* is a strong subjective (Riloff and Wiebe, 2003).
23*	Weak subjective	{true, false}	*w* is in Riloff and Wiebe's (2003) list of weak subjectives (e.g., *noisy*).
24*	Weak subjective in context	{true, false}	One/two word(s) around *w* is a weak subjective (Riloff and Wiebe, 2003).
25	Polarity	{+, –, both, ...}	The polarity of *w* according to Riloff and Wiebe (2003), e.g., *praising* is positive.
26*	Positive word	{true, false}	*w* is in Liu et al.'s (2005) list of positive words (e.g., *excel*).
27*	Positive word in context	{true, false}	One/two word(s) around *w* is positive (Liu et al., 2005).
28*	Negative word	{true, false}	*w* is in Liu et al.'s (2005) list of negative words (e.g., *terrible*).
29*	Negative word in context	{true, false}	One/two word(s) around *w* is negative (Liu et al., 2005).
30*	Grammatical relation	{root, subj, ...}	Whether *w* is the subject, object, root, etc. of its sentence.
31	Bias lexicon	{true, false}	*w* has been observed in NPOV edits (e.g., *nationalist*).
32*	Collaborative feature	<numeric>	Number of times that *w* was NPOV-edited in the article's prior history / frequency of *w*.

Table 3: Features used by the bias detector. The star (*) shows the most contributing features.

3. **Sentiment baseline**. Logistic regression model that only uses the features based on Liu et al.'s (2005) lexicons of positive and negative words (i.e., features 26–29).

4. **Subjectivity baseline**. Logistic regression model that only uses the features based on Riloff and Wiebe's (2003) lexicon of subjective words (i.e., features 21–25).

5. **Wikipedia baseline**. Selects as biased the words that appear in Wikipedia's list of words to avoid (Wikipedia, 2013a).

These baselines assessed the difficulty of the task, as well as the extent to which traditional sentiment-analysis and subjectivity features would suffice to detect biased language.

3.3 Results and Discussion

To measure performance, we used accuracy defined as:

$$\frac{\text{Number of sentences with the correctly predicted biased word}}{\text{Total number of sentences}}$$

The results are shown in Table 4. As explained earlier, we evaluated all the models by outputting as biased either the highest ranked word or the two or three highest ranked words. These correspond to the TOP1, TOP2 and TOP3 columns, respectively. The TOP3 score increases to 59%. A tool that highlights up to three words to be revised would simplify the editors' job and decrease significantly the time required to revise.

Our model outperforms all five baselines by a large margin, showing the importance of considering a wide range of features. Wikipedia's list of words to avoid falls very short on recall. Fea-

System	TOP1	TOP2	TOP3
Baseline 1: Random	2.18	7.83	9.13
Baseline 2: Role	15.65	20.43	25.65
Baseline 3: Sentiment	14.78	22.61	27.83
Baseline 4: Subjectivity	16.52	25.22	33.91
Baseline 5: Wikipedia	10.00	10.00	10.00
Our system	**34.35**	**46.52**	**58.70**
Humans (AMT)	37.39	50.00	59.13

Table 4: Accuracy (%) of the bias detector on the test set.

tures that contribute the most to the model's performance (in a feature ablation study on the dev set) are highlighted with a star (*) in Table 3. In addition to showing the importance of linguistic cues for different classes of bias, the ablation study highlights the role of contextual features. The bias lexicon does not seem to help much, suggesting that it is overfit to the training data.

An error analysis shows that our system makes acceptable errors in that words wrongly predicted as bias-inducing may well introduce bias in a different context. In (16), the system picked *eschew*, whereas *orthodox* would have been the correct choice according to the gold edit. Note that both the sentiment and the subjectivity lexicons list *eschew* as a negative word. The bias type that poses the greatest challenge to the system are terms that are one-sided or loaded in a particular topic, such as *orthodox* in this example.

(16) a. Some Christians eschew **orthodox** theology; such as the Unitarians, Socinian, [...]
b. Some Christians eschew **mainstream trinitarian** theology; such as the Unitarians, Socinian, [...]

The last row in Table 4 lists the performance of humans on the same task, presented in the next section.

4 Human Perception of Biased Language

Is it difficult for humans to find the word in a sentence that induces bias, given the subtle, often implicit biases in Wikipedia. We used Amazon Mechanical Turk[7] (AMT) to elicit annotations from humans for the same 230 sentences from the test set that we used to evaluate the bias detector in Section 3.3. The goal of this annotation was twofold: to compare the performance of our bias detector against a human baseline, and to assess the difficulty of this task for humans. While AMT labelers are not trained Wikipedia editors, understanding how difficult these cases are for untrained labelers is an important baseline.

4.1 Task

Our HIT (Human Intelligence Task) was called "Find the biased word!". We kept the task description succinct. Turkers were shown Wikipedia's definition of a "biased statement" and two example sentences that illustrated the two types of bias, framing and epistemological. In each HIT, annotators saw 10 sentences, one after another, and each one followed by a text box entitled "Word introducing bias." For each sentence, they were asked to type in the text box the word that caused the statement to be biased. They were only allowed to enter a single word.

Before the 10 sentences, turkers were asked to list the languages they spoke as well as their primary language in primary school. This was English in all the cases. In addition, we included a probe question in the form of a paraphrasing task: annotators were given a sentence and two paraphrases (a correct and a bad one) to choose from. The goal of this probe question was to discard annotators who were not paying attention or did not have a sufficient command of English. This simple test was shown to be effective in verifying and eliciting linguistic attentiveness (Munro et al., 2010). This was especially important in our case as we were interested in using the human annotations as an oracle. At the end of the task, participants were given the option to provide additional feedback.

We split the 230 sentences into 23 sets of 10 sentences, and asked for 10 annotations of each set. Each approved HIT was rewarded with $0.30.

4.2 Results and Discussion

On average, it took turkers about four minutes to complete each HIT. The feedback that we got from some of them confirmed our hypothesis that finding the bias source is difficult: "Some of the 'biases' seemed very slight if existent at all," "This was a lot harder than I thought it would be... Interesting though!".

We postprocessed the answers ignoring case, punctuation signs, and spelling errors. To ensure an answer quality as high as possible, we only kept those turkers who answered attentively by applying two filters: we only accepted answers that matched a valid word from the sentence, and we discarded answers from participants who did not

[7] http://www.mturk.com

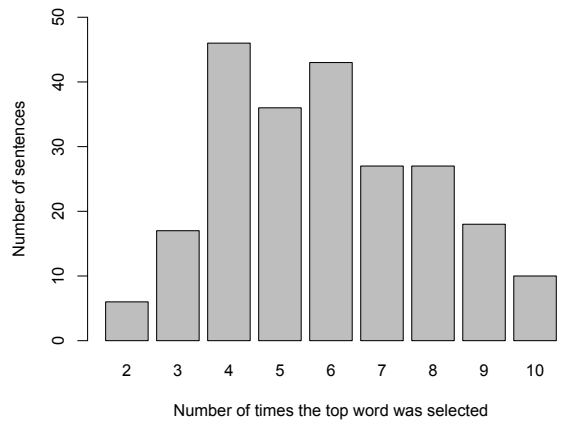

Figure 1: Distribution of the number of turkers who selected the top word (i.e., the word selected by the majority of turkers).

pass the paraphrasing task—there were six such cases. These filters provided us with confidence in the turkers' answers as a fair standard of comparison.

Overall, humans correctly identified the biased word 30% of the time. For each sentence, we ranked the words according to the number of turkers (out of 10) who selected them and, like we did for the automated system, we assessed performance when considering only the top word (TOP1), the top 2 words (TOP2), and the top 3 words (TOP3). The last row of Table 4 reports the results. Only 37.39% of the majority answers coincided with the gold label, slightly higher than our system's accuracy. The fact that the human answers are very close to the results of our system reflects the difficulty of the task. Biases in reference works can be very subtle and go unnoticed by humans; automated systems could thus be extremely helpful.

As a measure of inter-rater reliability, we computed pairwise agreement. The turkers agreed 40.73% of the time, compared to the 5.1% chance agreement that would be achieved if raters had randomly selected a word for each sentence. Figure 1 plots the number of times the top word of each sentence was selected. The bulk of the sentences only obtained between four and six answers for the same word.

There is a good amount of overlap ($\sim$34%) between the correct answers predicted by our system and those from humans. Much like the automated system, humans also have the hardest time identifying words that are one-sided or controversial to a specific topic. They also picked *eschew* for (16) instead of *orthodox*. Compared to the system, they do better in detecting bias-inducing intensifiers, and about the same with epistemological bias.

5 Related Work

The work in this paper builds upon prior work on subjectivity detection (Wiebe et al., 2004; Lin et al., 2011; Conrad et al., 2012) and stance recognition (Yano et al., 2010; Somasundaran and Wiebe, 2010; Park et al., 2011), but applied to the genre of reference works such as Wikipedia. Unlike the blogs, online debates and opinion pieces which have been the major focus of previous work, bias in reference works is undesirable. As a result, the expression of bias is more implicit, making it harder to detect by both computers and humans. Of the two classes of bias that we uncover, *framing bias* is indeed strongly linked to subjectivity, but *epistemological bias* is not. In this respect, our research is comparable to Greene and Resnik's (2009) work on identifying *implicit* sentiment or perspective in journalistic texts, based on semantico-syntactic choices.

Given that the data that we use is not supposed to be opinionated, our task consists in detecting (implicit) bias instead of classifying into side A or B documents about a controversial topic like ObamaCare (Conrad et al., 2012) or the Israeli-Palestinian conflict (Lin et al., 2006; Greene and Resnik, 2009). Our model detects whether all the relevant perspectives are fairly represented by identifying statements that are one-sided. To this end, the features based on subjectivity and sentiment lexicons turn out to be helpful, and incorporating more features for stance detection is an important direction for future work.

Other aspects of Wikipedia structure have been used for other NLP applications. The Wikipedia revision history has been used for spelling correction, text summarization (Nelken and Yamangil, 2008), lexical simplification (Yatskar et al., 2010), paraphrasing (Max and Wisniewski, 2010), and textual entailment (Zanzotto and Pennacchiotti, 2010). Ganter and Strube (2009) have used Wikipedia's weasel-word tags to train a hedge detector. Callahan and Herring (2011) have examined cultural bias based on Wikipedia's NPOV policy.

6 Conclusions

Our study of bias in Wikipedia has implications for linguistic theory and computational linguistics. We show that bias in reference works falls broadly into two classes, framing and epistemological. The cues to framing bias are more explicit and are linked to the literature on subjectivity; cues to epistemological bias are subtle and implicit, linked to presuppositions and entailments in the text. Epistemological bias has not received much attention since it does not play a major role in overtly opinionated texts, the focus of much research on stance recognition. However, our logistic regression model reveals that epistemological and other features can usefully augment the traditional sentiment and subjectivity features for addressing the difficult task of identifying the bias-inducing word in a biased sentence.

Identifying the bias-inducing word is a challenging task even for humans. Our linguistically-informed model performs nearly as well as humans tested on the same task. Given the subtlety of some of these biases, an automated system that highlights one or more potentially biased words would provide a helpful tool for editors of reference works and news reports, not only making them aware of unnoticed biases but also saving them hours of time. Future work could investigate the incorporation of syntactic features or further features from the stance detection literature. Features from the literature on veridicality (de Marneffe et al., 2012) could be informative of the writer's commitment to the truth of the events described, and document-level features could help assess the extent to which the article provides a balanced account of all the facts and points of view.

Finally, the NPOV corpus and the bias lexicon that we release as part of this research could prove useful in other bias related tasks.

Acknowledgments

We greatly appreciate the support of Jean Wu in running our task on Amazon Mechanical Turk, and all the Amazon Turkers who participated. We benefited from comments by Valentin Spitkovsky on a previous draft and from the helpful suggestions of the anonymous reviewers. The first author was supported by a Beatriu de Pinós postdoctoral scholarship (2010 BP-A 00149) from Generalitat de Catalunya. The second author was supported by NSF IIS-1016909. The last author was supported by the Center for Advanced Study in the Behavioral Sciences at Stanford.

References

Pranav Anand, Marilyn Walker, Rob Abbott, Jean E. Fox Tree, Robeson Bowmani, and Michael Minor. 2011. Cats rule and dogs drool!: Classifying stance in online debate. In *Proceedings of ACL-HLT 2011 Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, pages 1–9.

Jonathan Berant, Ido Dagan, Meni Adler, and Jacob Goldberger. 2012. Efficient tree-based approximation for entailment graph learning. In *Proceedings of ACL 2012*, pages 117–125.

Ewa Callahan and Susan C. Herring. 2011. Cultural bias in Wikipedia articles about famous persons. *Journal of the American Society for Information Science and Technology*, 62(10):1899–1915.

Eunsol Choi, Chenhao Tan, Lillian Lee, Cristian Danescu-Niculescu-Mizil, and Jennifer Spindel. 2012. Hedge detection as a lens on framing in the GMO debates: a position paper. In *Proceedings of the ACL-2012 Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics*, pages 70–79.

Alexander Conrad, Janyce Wiebe, and Rebecca Hwa. 2012. Recognizing arguing subjectivity and argument tags. In *Proceedings of ACL-2012 Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics*, pages 80–88.

Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC 2006*.

Marie-Catherine de Marneffe, Christopher D. Manning, and Christopher Potts. 2012. Did it happen? The pragmatic complexity of veridicality assessment. *Computational Linguistics*, 38(2):301–333.

Robert M. Entman. 2007. Framing bias: Media in the distribution of power. *Journal of Communication*, 57(1):163–173.

Viola Ganter and Michael Strube. 2009. Finding hedges by chasing weasels: Hedge detection using Wikipedia tags and shallow linguistic features. In *Proceedings of ACL-IJCNLP 2009*, pages 173–176.

Stephan Greene and Philip Resnik. 2009. More than words: Syntactic packaging and implicit sentiment. In *Proceedings of NAACL-HLT 2009*, pages 503–511.

Joan B. Hooper. 1975. On assertive predicates. In J. Kimball, editor, *Syntax and Semantics*, volume 4, pages 91–124. Academic Press, New York.

Ken Hyland. 2005. *Metadiscourse: Exploring Interaction in Writing*. Continuum, London and New York.

Lauri Karttunen. 1971. Implicative verbs. *Language*, 47(2):340–358.

Paul Kiparsky and Carol Kiparsky. 1970. Fact. In M. Bierwisch and K. E. Heidolph, editors, *Progress in Linguistics*, pages 143–173. Mouton, The Hague.

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 CoNLL 2006*, pages 109–116.

Chenghua Lin, Yulan He, and Richard Everson. 2011. Sentence subjectivity detection with weakly-supervised learning. In *Proceedings of AFNLP 2011*, pages 1153–1161.

Bing Liu, Minqing Hu, and Junsheng Cheng. 2005. Opinion Observer: analyzing and comparing opinions on the Web. In *Proceedings of WWW 2005*, pages 342–351.

Aurélien Max and Guillaume Wisniewski. 2010. Mining naturally-occurring corrections and paraphrases from Wikipedia's revision history. In *Proceedings of LREC 2010*, pages 3143–3148.

Robert Munro, Steven Bethard, Victor Kuperman, Vicky Tzuyin Lai, Robin Melnick, Christopher Potts, Tyler Schnoebelen, and Harry Tily. 2010. Crowdsourcing and language studies: the new generation of linguistic data. In *Proceedings of the NAACL-HLT 2010 Workshop on Creating Speech and Language Data With Amazons Mechanical Turk*, pages 122–130.

Rani Nelken and Elif Yamangil. 2008. Mining Wikipedias article revision history for training Computational Linguistics algorithms. In *Proceedings of the 1st AAAI Workshop on Wikipedia and Artificial Intelligence*.

Souneil Park, KyungSoon Lee, and Junehwa Song. 2011. Contrasting opposing views of news articles on contentious issues. In *Proceedings of ACL 2011*, pages 340–349.

Ellen Riloff and Janyce Wiebe. 2003. Learning extraction patterns for subjective expressions. In *Proceedings of EMNLP 2003*, pages 105–112.

Swapna Somasundaran and Janyce Wiebe. 2010. Recognizing stances in ideological on-line debates. In *Proceedings of the NAACL-HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 116–124.

Peter D. Turney. 2002. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of ACL 2002*, pages 417–424.

Janyce Wiebe, Theresa Wilson, Rebecca Bruce, Matthew Bell, and Melanie Martin. 2004. Learning subjective language. *Computational Linguistics*, 30(3):277–308.

Wikipedia. 2013a. Wikipedia: Manual of style / Words to watch. `http://en.wikipedia.org/wiki/Wikipedia:Words_to_avoid`. [Retrieved February 5, 2013].

Wikipedia. 2013b. Wikipedia: Neutral point of view. `http:http://en.wikipedia.org/wiki/Wikipedia:Neutral_point_of_view`. [Retrieved February 5, 2013].

Tae Yano, Philip Resnik, and Noah A. Smith. 2010. Shedding (a thousand points of) light on biased language. In *Proceedings of the NAACL-HLT 2010 Workshop on Creating Speech and Language Data With Amazons Mechanical Turk*, pages 152–158.

Mark Yatskar, Bo Pang, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2010. For the sake of simplicity: Unsupervised extraction of lexical simplifications from Wikipedia. In *Proceedings of NAACL-HLT 2010*, pages 365–368.

Fabio M. Zanzotto and Marco Pennacchiotti. 2010. Expanding textual entailment corpora from Wikipedia using co-training. In *Proceedings of the 2nd Coling Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources*, pages 28–36.