Biases We Live By

Anders Søgaard

Bio

Anders Søgaard, b. 1981, is Professor with Special Duties in NLP and Machine Learning at University of Copenhagen. He holds an ERC Starting Grant and has won several research prizes and best paper awards at top computer science venues.

Abstract

Moderne teknologier er brugervenlige, fordi de har indbygget en "bias" (partiskhed), der forudser og tilpasser sig vores behov, men sommetider diskriminerer en sådan tilpasningsmekanisme også mellem forskellige brugergrupper og favoriserer visse demografier over andre. Artiklen diskuterer hvorledes sådanne "biases" opererer indenfor søgemaskiner og sprogteknologisk software.

Anders Søgaard, Professor
Center for Language Technology,
University of Copenhagen
soegaard@hum.ku.dk

Introduction

The economist Ha-Joon Chang famously argued that the invention of the washing machine had greater impact on society than the invention of the Internet. The reason is simple, namely, that the washing machine-along with the invention of other household conveniences-eventually led to women's entrance into the job market. Nevertheless, the Internet means the world to most of us-in our jobs, when we travel, when we plan our weekends, or order food from a nearby restaurant.

The crux here is the rapid availability of information. The Internet makes information available to us in a natural, efficient way. Search engines such as Google try to optimize the quality of search, so that the information sought is always found on the highest-ranked website returned. Google also makes searches accessible to as many as possible, in a natural way, introducing the question-answering features of Google Search and Voice Search, for example. Combining the two, you can now ask your Google Search App questions such as "When was Martin Luther King born?" or "What is the time in Atlanta?" and have your mobile phone read aloud the answer. Search quality, question answering, and speech recognition all rely on basic language technologies, such as the ability to automatically disambiguate words (e.g., inferring whether *beat* is a noun or a verb in a particular search query). Below I briefly discuss some of the biases in basic language technologies, and how they might lead to different demographic groups having

very different experiences with the Internet. Moreover, question answering typically relies on facts harvested from Wikipedia, introducing more potential biases.

Bias

The word *bias* is often said to be of Greek origin, from the word *epikarsios*, meaning 'oblique'. When people have partial perspectives, we often say that people are biased. A biased search is one with only partial access to information, or one that results in rankings that systematically favor certain results.

Part of the success of Internet search engines is that they are heavily biased. A search engine is likely to return websites that are popular. Also, most search engines are personalized, and hence biased in favor of websites similar to those you visited before. Both these features lead to user-friendly search experiences, and neither of the biases systematically favors one users over another. In other words, neither of these biases are political.

Other biases *are* political. When the search engine Bing operates in China, state authorities filter the search results. Major search engines also try to avoid returning links to child pornography or sites in conflict with anti-terror laws, for example.

The bias introduced by language technologies is subtler, yet still of political interest. Search quality differs between different demographics-for reasons I will address below-with significantly worse results for young women or African Americans than for older White men. Disambiguation, syntactic analysis, and speech recognition all work best for older White men, and the search results returned for this group are thus, statistically speaking, more accurate. Below I refer to some research papers that show evidence for this bias and attempt to explain how the bias came about.

We might speculate that the use of Wikipedia as a knowledge base for question answering reinforces the gender and racial biases, since it is well known that most authors of Wikipedia entries are White men. However, a recent study (Wagner et al., 2015) suggests this is not the case. We briefly review their findings.

Bias in language technologies

It is perhaps hard to imagine that small pieces of software that decide whether *break* is a noun or a verb can introduce important demographic biases. As Hovy and Søgaard have shown (2015), such software, it seems, is much better at analyzing texts if authored by White men in the 45+ age range. The reason for this bias is simple, once we understand how this software comes about. But let us first establish that there are gender and age differences in how we use language.

It is well known that men and women differ in their use of certain syntactic categories. Men use more numbers. Men are more than 30% more likely to use numbers, for example. Women use more pronouns, while men use compound nouns more often. In recent work (Johannsen et al., 2015) we presented evidence suggesting that, across a range of Indo-European languages, women systematically coordinate verb phrases more often than men. This holds not only in English, but in *all* the languages studied; every time men coordinate verb phrases 5 times, women do it six times.

Here's how language technologies result in bias: software that analyzes words in texts, determining their syntactic category or grammatical function, is based on algorithms for finding patterns and generalizations across data, inducing models from manually-annotated text collections. In other words, language technology generalizes across hundreds or thousands of sentences analyzed by hand by professional linguists. Our algorithms induce knowledge of language and grammar from the examples, enabling us to analyze new, unseen sentences.

If our algorithms are run on English sentences, we learn models for English. If run on Basque, we learn models for Basque. However, for practical and historical reasons, our manually-annotated text collections are typically not very representative of languages. When language technology got off the ground in the late 80s and early 90s, most digital text was newswire. So this was the text that researchers decided to annotate. Thousands of newspaper articles have been annotated for English, German, Spanish, but also for Arabic and Chinese. This of course means that our models are better for newswire than

for other types of language, for example poetry or spoken language.

Also, journalists are not a representative sample of a population. Journalists are-in the US and in Europe-typically White men over the age of 45. The readers of newspapers such as the *Wall Street Journal* (used for most English linguistic resources) and *Frankfurter Rundschau* (used for many German resources), for example, also tend to be 45+ White men. Knowing this, it is not a surprise that our language technologies, induced from annotated texts for and by this group, perform much better on texts written by their peer demographic. If the text has been trained on texts with more numbers and compound nouns, our models will expect to see numbers and compound nouns and will need less evidence for such constructions, leading to false positives. Hovy and Søgaard (2015) evaluated state-of-the-art syntactic analyzers from different demographic groups in the US and Germany, showing that analyzers perform better on the texts by older men. They demonstrated that this was not just due to unseen words in adolescent language, but also due to grammatical differences between the age groups. The result, then, means older White men get better Internet searches and better voice recognition than the rest of us.

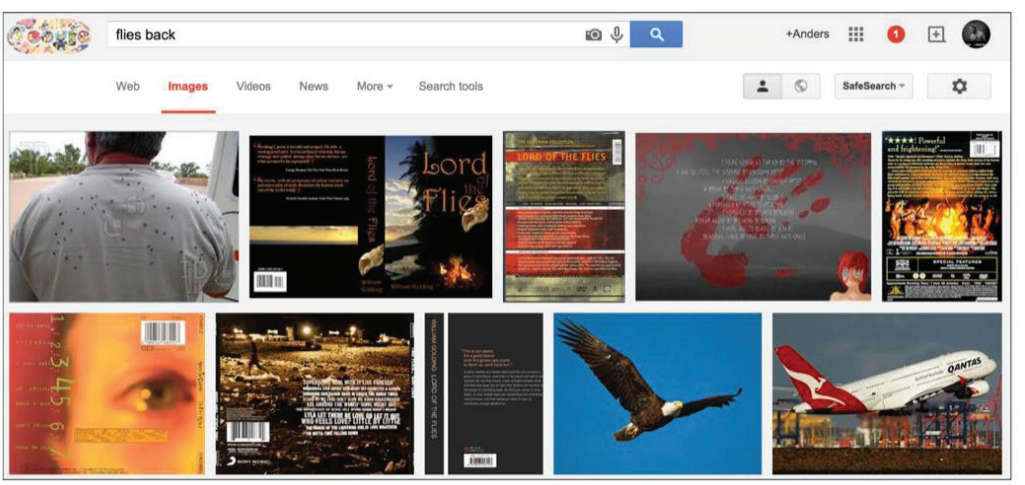
To comprehend this, consider the (surprisingly frequent) search query: *flies back*. This search query returns several things: back covers of Alice in Chains album *Jar of Flies*, as well as the of the book *Lord of the Flies*; stories of men covered in flies; and stories of eagles and where they fly. In the case of album and book covers, the query is interpreted as a proper noun followed by a common noun. In the second case, which returns men covered in flies, the words are recognized as two consecutive common nouns. In the third case, they are read as a verb and an adverb. The search query is truly ambiguous, but if we add more contexts, the two words are disambiguated (e.g., *flies back cover, flies on back,* or *flies back to*). The ability to automatically draw such inferences considerably improves the search results (Ganchev et al., 2013).

Syntactic analysis also improves speech recognition. Consider, for example, the pronunciation of *bag* and *back*, in many cases the same. If we hear the sound |*'bæg*| after an intransitive verb such as *flies*, we know the next word cannot be a noun or a verb. This rules out the word *bag*, and we recognize the sound as the word *back*.

Finally, biases in language technology are not limited to gender and age. Jørgensen et al. (2015) quantify the performance drops seen with automated analysis of African American Vernacular English (AAVE). Here performance drops are significant, even more than those seen across gender and age groups.

Bias in Wikipedia

Wikipedia is an encyclopedic portrait of our world, written, almost exclusively, by young well-educated men. Since Wikipedia is the biggest publicly-available source of semi-structured knowledge and often used as a knowledge base for question answer-



ing, this bias could potentially percolate to Internet search.

Wagner et al. (2015) recently explored how Wikipedia is gender-biased, studying coverage biases, structural biases, and content biases. They used existing knowledge bases to extract Wikipedia's entries about notable persons. Coverage bias is the question of whether Wikipedia covers the same proportion of notable women as they do with men. Structural bias relates to the structure of the Wikipedia citation network, while content bias looks at differences in how women and men are described on Wikipedia. Wagner et al. (2015) found that across several language versions, Wikipedia covers female notable persons slightly better than male notable persons, relative to their knowledge bases. However, the knowledge bases may of course be severely biased, something the authors do not discuss. Unsurprisingly, the authors find that entries about women more often link to entries about women, and the same for men. The authors also explore other biases, for example, how entries about notable women are more likely to mention biographical details about their family.

The structural and content biases are not important for how Wikipedia is used in question answering, however. Only coverage bias, which Wagner et al. (2015) claim is non-existent, is important. Again, the coverage bias they estimate is relative to knowledge bases that may also be biased. Finally, note that gender bias is only one concern people have raised about Wikipedia's coverage. Others have worried about the poor coverage of Black history, for example.

Bias - good or bad?

Search engines work precisely because they are biased. When I search restaurants, for instance, Google returns Copenhagen restaurants rather than restaurants in Atlanta. When I search Michael Jordan, my top hit is about the researcher, not the basketball player. Information is abundant, and if we did not bias our search engines, we would have to compose one-page search queries to find what we are looking for. Also, the search would likely be much slower.

Biased search engines are biased (oblique) in the sense that they make assumptions about the user. But assumptions may be more or less appropriate. There are parts of search engines that assume users are 45+ White men, which is clearly not always appropriate. It was maybe more appropriate in the early 90s, when researchers started working on search engines and language technologies. But these days, the Internet, statistically speaking, belongs to Black teenage girls, and search tools are no longer adequate.

Washing machines (along with other household commodities) changed life in the Western world, eventually enabling women to enter the job market. Washing machines are also biased, making assumptions about how much laundry we produce and how we want it washed. In other words, bias is not good or bad. Bias is necessary, comprising a model of users and their usage. Some models are good, some are bad, but it is not about whether they are biased or not.

The Internet has made our life much easier, in the Western world, but unlike washing machines, the usefulness of the Internet depends a lot on whether you are in North America or in India. In North America, most people speak English, Spanish, or French-all major languages with good language technology support for Internet search, including question answering and voice search. India has twenty-two official languages and 1650 dialects, few of which have any language technology support whatsoever. In addition, roughly speaking, language technology seems to favor the wealthy. If the Internet is to be a driver of large-scale societal change, we need to make sure that the Internet is (at least) equally useful to the under-resourced across the globe as it is to university professors, golf-playing bankers and retired lawyers in the developed world.

This work is funded by the ERC Starting Grant LOWLANDS No. 313695.

References

Ganchev, K, Hall, K, McDonald, R & Petrov, S (2013). Using Search-Logs to Improve Query Tagging. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, (pp. 238-242), Jeju: Republic of Korea.

Hovy, D & Søgaard, A (2015). Tagging Performance Correlates with Author Age. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint*

Conference on Natural Language Processing (Short Papers), (pp. 483-488), Beijing China.

Johannsen, A, Alonso, HM & Søgaard, A (2015). Any-Language Frame Semantic Parsing. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, (pp. 2062-2066), Lisbon, Portugal.

Jørgensen, AK, Hovy, D & Søgaard, A (2015). Challenges of Studying and Processing Dialects in Social Media. *Proceedings of the ACL 2015 Workshop on Noisy User-generated Text*, (pp. 9-18), Beijing, China, July 31, 2015.

Wagner, C, Garcia, D, Jadidi, M & Strohmaier, M (2015). It's a Man's Wikipedia? Assessing Gender Inequality in an Online Encyclopedia. *The International AAAI Conference on Web and Social Media* (ICWSM2015), Oxford, May 2015.