The Importance of Modeling Social Factors of Language: Theory and Practice

Dirk Hovy
Bocconi University
Via Sarfatti 25
20136 Milan, Italy
dirk.hovy@unibocconi.it

Diyi Yang
Georgia Institute of Technology
CODA Tech Square
Atlanta, GA, 30308
dyang888@gatech.edu

Abstract

Natural language processing (NLP) applications are now more powerful and ubiquitous than ever before. With rapidly developing (neural) models and ever-more available data, current NLP models have access to more information than any human speaker during their life. Still, it would be hard to argue that NLP models have reached human-level capacity. In this position paper, we argue that the reason for the current limitations is a focus on information *content* while ignoring language's *social factors*. We show that current NLP systems systematically break down when faced with interpreting the social factors of language. This limits applications to a subset of information-related tasks and prevents NLP from reaching human-level performance. At the same time, systems that incorporate even a minimum of social factors already show remarkable improvements. We formalize a taxonomy of seven social factors based on linguistic theory and exemplify current failures and emerging successes for each of them. We suggest that the NLP community address social factors to get closer to the goal of human-like language understanding.

1 Introduction

> "[T]he common misconception [is] that language use has primarily to do with words and what they mean. It doesn't. It has primarily to do with people and what *they* mean."
> Clark and Schober (1992)

Until the 1970s, economics assumed that individuals, markets, and firms always acted rationally, based on all the available information. This assumption allowed researchers to use linear models and worked well for several applications. However, it came at the cost of ignoring essential aspects of human decision making, which oversimplified an inherently complex matter in a way that limited possible insights and applications. The seminal work by Tversky and Kahneman (1973) showed that people would make irrational decisions, time and again, even with full information, and that simple models could not account for this behavior. By introducing the human factor into the equation, they opened up a new research field: behavioral economics.

Like economics in the mid-twentieth century, Natural Language Processing (NLP) still makes a limiting assumption: language is only about information, i.e., message content alone. This assumption makes it possible to model language statistically and works for several applications. However, it completely ignores the fact that people use language to achieve (social) goals; like economists before 1973, NLP researchers are oversimplifying an inherently complex matter in a way that limits its possible insights and applications. And like introducing behavior transformed economics, introducing social factors into NLP will similarly transform the field: it will open up new avenues of research, enable new insights and applications, and provide more performant, equitable tools.

The focus on information content is rooted in early research on quantifying text and making it usable for information retrieval. While it oversimplifies its subject matter, this focus has enabled many NLP applications, with increasing commercial success over the last few decades. The statistical revolution and introduction of machine learning in the late 1980s and deep learning in the last five years (Manning, 2015) has dramatically improved robustness and performance, and produced industrial-strength everyday applications like machine translation (Wu et al., 2016), search (Shen et al., 2014), and personal assistants (Serban et al., 2016; Radford et al., 2019). Recently, BERT (Devlin et al., 2019) and GPT-3 (Brown et al., 2020) seemingly picked up enough language behavior to produce natural-looking sentences that show pragmatic constraints and interact in dialogues. However, recent work has pointed out (Bender and Koller, 2020; Bisk et al., 2020) that language is more than just words strung together: it has a social function and relates to non-linguistic context. Nonetheless, current NLP systems still largely ignore the social aspect of language. Instead, they only pay attention to *what* is said, not to *who* says it, in what *context*, and for which *goals*.

We go further to argue that the simplifying focus on information content has effectively limited NLP to a narrow range of information-based applications. Consequently, NLP systems struggle with applications related to pragmatics and interaction, or when "what is said is not what is meant," e.g., sarcasm, irony, deception, and any other situation that requires a "social" interpretation (Abercrombie and Hovy, 2016). This approach is especially crucial for any system related to pragmatics, such as dialogue systems, machine translation (Mirkin and Meunier, 2015), text-to-speech, and mental healthcare tools (Benton et al., 2017). Examples include conversational agents' inconsistent personality in conducting dialogues with humans (Cercas Curry et al., 2020), the failure of machine translation systems in generating culturally appropriate and polite outputs (Jones and Irvine, 2013; Matusov, 2019; Vanmassenhove et al., 2019), or the general struggles of current systems with social intelligence (Cercas Curry and Rieser, 2018).

Ultimately, the goal of NLP is to process language at a human level. However, NLP's current approach—ignoring social factors—prevents us from reaching human-level competence and performance because language is more than just information content. Unless we start paying attention to the social factors of language, we are artificially limiting NLP's potential as a field and the applications we can develop, including the performance of the applications that exist today.

We want to be clear that the idea of language as a social construct is itself nothing new: linguistics and philosophy have long modeled it this way (Wittgenstein, 2010; Eckert, 2012, inter alia). However, as we are reaching a point where this idea can become implemented in systems, it is a message that bears repeating in the NLP community (see also Hovy (2018) and Flek (2020) for similar points, as well as Nguyen et al. (2016) for an overview of the closely related issue of computational sociolinguistics). There have indeed been ongoing and emerging efforts to overcome these limitations. Over the last ten years, research interest in social factors and social context has increased, as shown in Figure 1. Here, we counted the number of accepted papers for the track of computational social science and social media, sentiment analysis, discourse and pragmatics, and their sum at the ACL conference per year, and visualized the overall trend [1]. However, to further highlight and formalize these social factors in language and their use in NLP, we propose a set of seven social factors, explain why they are needed, and show encouraging evidence of approaches that have used them. We hope that this work can inspire more research into the social factors of language in NLP, and push the boundary of what we can achieve as a research field.

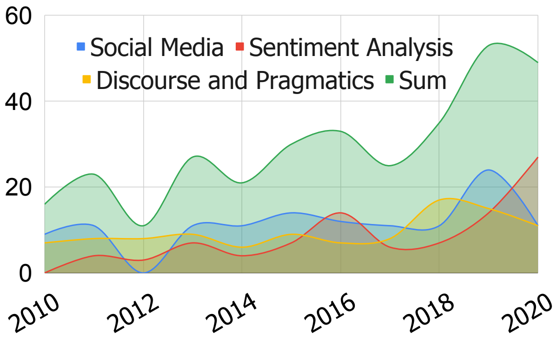

Figure 1: Trend of interest in social factors in NLP papers, using ACL as an example

Contributions We formalize the notion of social factors via two linguistic theories: systemic functional linguistics (Halliday and Matthiessen, 2013, SFL) and the Cooperative Principle (Grice, 1975). We build on these frameworks to provide a taxonomy of seven increasingly complex social factors that help tease out the limitations of NLP models. These seven factors are: 1) **speaker** and 2) **receiver**, 3) **social relations**, 4) **context**, 5) **social norms**, 6) **culture and ideology**, and 7) **communicative goals**. For each factor, we explain why it presents an obstacle to current information-based approaches and show work that has started to address them.

2 Taxonomy of Social Factors

Systemic functional linguistics (SFL) (Halliday and Matthiessen, 2013), studies precisely this re-

[1] https://public.flourish.studio/visualisation/2431551/

lationship between language and its functions in social settings. It gives us a sense of the different language areas that, instead of formal factors like syntax and semantics, rely on social factors for interpretation. By detailing those factors, we can understand what is missing in current NLP approaches, and how to incorporate them into our systems to go beyond information content.

However, SFL alone can not explain why "what is said is not what is meant." For that, we borrow from Grice (1975), who laid out four maxims that govern effective communication in social situations. These four maxims are those of *Quality* ("Make your contribution true, do not lie or make unsupported claims"), *Quantity* ("Make your contribution as informative as is required (but not more informative)"), *Relevance* ("Make your contribution relevant"), and *Manner* ("Be brief and orderly and avoid obscurity of expression and ambiguity"). Together, these maxims are known as the *Cooperative principle*, and govern successful conversations, as long as all conversational partners adhere to them.

However, we can also deliberately break selected maxims, for example, for comical effect, sarcasm, politeness, when we playact, or outright lie (i.e., saying things that are not true, not relevant, or obtuse). If this violation is apparent, the conversational partner can use the resulting inconsistency to construct an alternative meaning. E.g., inferring that "Take your time, I *love* waiting for you" violates the maxim of quality and is probably not true lets us assume sarcasm. Gricean maxims and their selective violations can explain *why* "what is said is not what is meant." This inference process is called *conversational implicature*, and can help explain why NLP applications struggle with tasks such as sarcasm detection or entailment. Some previous works have consequently used them to evaluate the quality of NLP systems (Jwalapuram, 2017; Qwaider et al., 2017).

Building upon these two frameworks, we lay out a set of seven social factors that NLP systems need to be aware of to overcome current limitations (see Figure 2). We cover SPEAKER characteristics (Section 2.1), RECEIVER characteristics (Section 2.2), SOCIAL RELATIONS (Section 2.3), CONTEXT (Section 2.4), SOCIAL NORMS (Section 2.5), CULTURE AND IDEOLOGY (Section 2.6), and COMMUNICATIVE GOALS (Section 2.7). We first outline each factor and its relation to SFL and the cooperation principle and then discuss the associated limitations for current NLP systems, as well as existing approaches that address these factors.

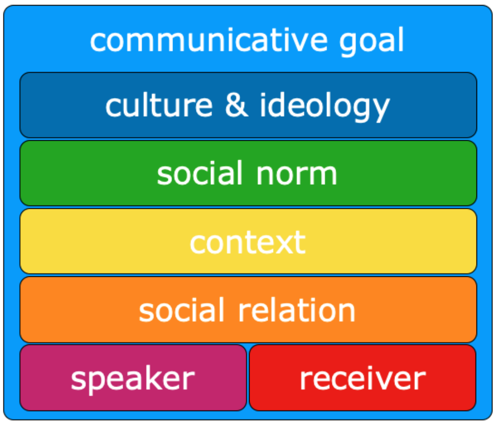

Figure 2: Taxonomy of social factors

Note that the seven social factors in this taxonomy are *not* mutually exclusive. Most language use can be categorized according to multiple factors, such as the use of goal and norm.

2.1 Speaker

An individual or agent uses language for different social goals, such as constructing their identity. Characteristics of speakers include age, gender, ethnicity, social class, dialect, etc. A speaker determines the speech act, text, tone, language style, and consciously encoded personal signatures of an utterance. Certain speaker attributes are expected to be *consistent* or unchanged across different scenarios, such as basic demographics and personality traits. Other can *vary* according to situation, such as tone and style. In both cases, the speaker has a certain amount of *agency* over the expression of some of these attributes, but will be unaware of others. In sociolinguistics, this hierarchy is called *saliency*, ranging from obvious to all speakers (e.g., "howdy" for Texans) to apparent only to speakers of the variety (e.g., when to unround a vowel or not), or only to researchers (e.g., syntactic inversion) (Silverstein, 2003). Successful speaker models should thus use the cooperative principle as a set of constraints and know when to break them for effect.

Applications Failing to consider speaker characteristics might result in inaccurate models, e.g., the message of a 20-year-old German female reading like it was from a 75-year-old American male after translation (Hovy et al., 2020). This effect is a big issue for any text generation, where the lack of speaker personality can create incongruous responses in conversational agents. Despite conversational agents' recent successes (Ritter et al., 2011; Banchs and Li, 2012; Serban et al., 2016), their lack of a consistent personality is still one of the common issues in using data-driven approaches. The main reason is that these models are often trained over conversations by different people, averaging and thereby virtually ignoring individual speakers' personalities (Li et al., 2016; Wei et al., 2017; Zhang et al., 2018; Wu et al., 2021). There have not been many attempts to make NLP systems more robust to language variation across speakers (Yang and Eisenstein, 2017), though attempts at creating personalized language technologies exist in information retrieval (Shen et al., 2005), recommender systems (Basilico and Hofmann, 2004), machine translation (Mirkin and Meunier, 2015), and language modeling (Federico, 1996). Meanwhile, various approaches have shown the positive impact of incorporating speaker characteristics into NLP applications, either as explicit features (Volkova et al., 2013), through conditional embeddings (Hovy, 2015; Lynn et al., 2017), or via neural models for multi-task learning (Benton et al., 2017; Li et al., 2018). By accounting for a speaker's specific demographic attributes, models achieve better performance in a variety of tasks, such as sentiment analysis, user attributes, part-of-speech tagging, and response generation (Wu et al., 2021). Rashkin et al. (2016) showed the value of modelling speaker perspective to discover opinions or biases in the way things are expressed. Hovy (2016) showed that demographically-conditioned generated text also is more convincing.

2.2 Receiver

Audiences that receive text from a speaker are made up of receivers, depending on the situation and medium. The number of receivers can vary substantially, ranging from zero (monologue) to one (dialogue), multiple (conversation), or massive (broadcast). Receivers may be known or unknown. For instance, in any given dialogue or conversation, the speaker knows the identity of the specific and fixed target or group to whom he/she is talking. However, when it comes to broadcasting or highly public spaces, receivers are often "imagined" by the speaker (Litt, 2012) and are potentially numerous and invisible. This imagined audience is a speaker's mental conceptualization of the people with whom he or she is communicating. This conceptualization of receiver characteristics influences the conversation: a speaker who calls on Newton's "Celestial Mechanics" to respond to a child's question "Where does the sun go at night?" has grossly misconceptualized the receiver characteristics in the situation.

Successful receiver models should thus use the cooperative principle as a set of constraints on what to expect from a counterpart. However, they should also assume that the receiver will perform conversational implicature when they notice a maxim violation. Right now, conversational agents tend to take any input as adhering to all maxims, so they are bad at recognizing sarcasm, irony, or overly polite forms (all of which violate the maxim of quality by saying things that are not true: you really *do* want another piece of cake).

Applications Spellchecking and stylistic models currently fail to consider receiver characteristics. For instance, when *writing to the president of a company* vs. *messaging your best friend*, the politeness levels and register differ substantially, but current large, pretrained models cannot deal with this difference effectively (for an exception, see Fu et al. (2020)). What is more, they can generate messages that are actively hurtful to receivers (Nozza et al., 2021). In other cases like hateful-content detection (Warner and Hirschberg, 2012), a message might be toxic to outsiders but perceived as appropriate among close friends (Sap et al., 2019a). This self-reference or joking use of slurs by a group of intimates might introduce significant noise to the automatic recognition of hate speech, causing existing classifiers to fail in many instances. Detecting such hateful or toxic speech online might require classifiers to take into account both content and receivers, as well as a broader context. Receiver differences markedly add to the complexity and difficulty in machine translation from, say, English to Korean. Korean speech has strict rules about politeness in language depending on who you are talking to; misusing these measures would be viewed as quite rude by native speakers of Korean (Kim and Lee, 2017).

2.3 Social Relation

The distance or relation between speaker and receiver matters. Examples of social relations include family, friendship, rival, ally, competitor, professional hierarchies, seniority, follower, and followee. One of the core communicative functions of language is to establish, modulate, and reproduce these social dynamics and social relations (Hymes, 1972). The interplay between speakers, receivers, and their relations introduces variations and flexibility into the resulting text. It also provides a shared background knowledge and context (this function of social relations has also influenced work on meaning frames by Fillmore (1982)). The incorporation of social relations is closely related to the consideration of speakers and receivers, but with different roles. In various social relations, we can flaunt the maxim of manner by being obscure, since much of the missing information will be filled in by shared knowledge.

Applications We could improve the detection of self-referential or joking use of hateful content with close friends if we could understand such social relations in the first place, similar to the context of response generation for different audiences. For the sentiment classification task, Yang and Eisenstein (2017) argue that models fail to leverage the tendency of socially proximate individuals (e.g., friends) to use language similarly. Ignoring this phenomenon of linguistic homophily usually means they suffer from limited accuracy. In practice, such social relations often can be reasonably inferred from text (Krishnan and Eisenstein, 2015; Iyyer et al., 2016; Rashid and Blanco, 2017; Rashid et al., 2020). They go a long way to explaining other socially motivated constructs, such as power imbalances or politeness, which in turn can also be inferred from dialogue (Prabhakaran et al., 2012; Danescu-Niculescu-Mizil et al., 2013a). Radfar et al. (2020) showed that including friendship relations in their hate-speech detection improved performance by up to 5%. Similarly, Del Tredici et al. (2019) showed that modeling the social graph of a user improves performance in sentiment analysis, as well as stance and hate speech detection. Incorporating user networks into geolocation substantially improves performance (Rahimi et al., 2018; Fornaciari and Hovy, 2019) and Dinan et al. (2020) show that the different roles of speaking-as, speaking-to, and speaking-about affect gender bias in NLP models.

Certain word choices or pronunciations might signal social class, status, or membership in a dialect group. Labov (1972) famously showed how realization of the /r/ sound in phrases like “fourth floor” was correlated with social hierarchy. In sociolinguistics (Trudgill, 2000), these distinguishing terms are called *shibboleths*, based on a story from the Old Testament in which pronouncing the word *shibboleth* a certain way decided whether a person was allowed to pass a checkpoint or was killed. Dialectal areas still play an important role, even in online communication (Hovy and Purschke, 2018), and identifying and integrating them can be vital for fairer NLP tools (Jørgensen et al., 2016; Blodgett et al., 2016; Dorn, 2019).

2.4 Context

Language-based communication usually takes place in a limited number of social contexts. These contexts reflect the detailed settings speakers and receivers are in, including (but not limited to) the language (e.g., English), domain (e.g., Twitter), occasion (e.g., presentation or discussion), and topic (e.g., work or life). As the “containers” or “holders” of communication (Yang, 2019, p. 20), (interpersonal) contexts set the specific boundaries for exchanging language. Prior research on dialogue (Schank and Abelson, 1975) accounted for (social) context as “scripts”, but framed it in terms of content rather than social factors.

Social context is related to the Gricean maxims of quantity and relevance, as it governs what is appropriate and required. Randomly (i.e., without context) saying “I have never smuggled live animals in my underwear” would probably raise some justified suspicion. In contrast, it is a perfectly acceptable response to the question, “Did you hide that parrot in your underpants?” (whether the question is appropriate is another matter).

Applications NLP models, by their nature, are usually unaware of the (extralinguistic) context. For instance, text or response generation may need to adaptively adjust to the social context of communication, rather than relying on background conversations from different communicators in different contexts. Models have mostly learned to relate words to other words. For instance, current machine translation models are trained on huge corpora of text. However, nuances in language often make it difficult to provide an accurate and direct translation from one social context to another. Studies show that current popular industrial MT systems and recent state-of-the-art academic MT models are significantly prone to gender-biased translation errors for all tested target languages (Stanovsky et al., 2019; Vanmassenhove et al., 2019; Hovy et al., 2020). There is hilarious content caused by translation fails (see #translationfail on Twitter), especially when it comes to the social context or cultural-specific nuances of language. Current text generation models also usually fail to account for social context, generating text that lacks nuance.

This factor is one of the most difficult ones to overcome, because 1) social context is almost always extralinguistic, and 2) the focus of NLP models has always been on learning applications based on text alone (amplified by the seeming ability of neural approaches to do so, see Collobert et al. (2011)). Some recent papers have commented on the artificial limitation of relying solely on text (Bender and Koller, 2020; Bisk et al., 2020), demonstrating how even large pretrained language models are essentially just mimicking people's language use, instead of *actual* use. Several works have shown, though, how incorporating non-textual information can improve performance, specifically in conjunction with images (Lazaridou et al., 2015; Caglayan et al., 2019). These approaches help various tasks, from concept learning to machine translation, and improve inherently multimodal applications such as scene descriptions and image labeling. However, even including more linguistic context (i.e., text beyond the current sentence) can drastically improve performance of text classification (Yang et al., 2016) and the detection of irony (Wallace et al., 2014) and sarcasm (Abercrombie and Hovy, 2016).[2]

2.5 Social Norm

Social norms refer to acceptable group conduct, shared understandings, or informal rules, representing speakers' and receivers' basic knowledge of what others do and what others think they should and should not do (Fehr and Fischbacher, 2004), such as dining etiquette, community norms on Reddit (Chandrasekharan et al., 2018), or hierarchical greetings. Norms are therefore closely related to the factors of relation (Section 2.3) and context (Section 2.4). For instance, greeting messages are usually full of positive words and phrases and rarely contain expressions carrying strong negative connotations. Product representatives are expected to communicate with customers in a professional manner rather than teasing or using slang and informal words. The scope of norms also include social commonsense about what is expected and "normal" in a given situation (Sap et al., 2019b), similar to scripts in Schank and Abelson (1975).

Social norms are related to the Gricean maxims of manner and quality: in some situations, it is very much expected to say too much and make unsupported claims, for example, when giving a laudatory speech or a eulogy; "*Good evening. Martin didn't stand out while he was alive. Now he is dead. Thank you.*" is not much of a speech.

Applications Social norms are subtle constructs that are not easy to define, so we still do not have many computational techniques to reliably quantify them, let alone assessing whether certain model behaviors should be rewarded or sanctioned (Anastassacos et al., 2020). Consequently, most NLP models still fail to recognize social norms (for an exception, see Forbes et al. (2020)).

Failing to measure social norms, and to detect the alignment between expected or unexpected behaviors and models' actual behaviors, can introduce severe damage and negatively impact society, especially as more conversational agents or chatbots have been developed and deployed for real-world applications, such as customer services, travel or flight reservation, or therapy. In 2016, Microsoft released its now infamous chatbot on Twitter: Tay[3]. Microsoft initially expected Tay's language patterns to resemble a 19-year old American girl, but the chatbot quickly transformed into a fountain of racist, sexist, and abusive slurs, by interacting with people espousing these views. A similar issue played out recently with a Korean chatbot.[4]

Sap et al. (2019a) showed that lack of awareness of social norms around taboo words led to annotation bias being integrated into the models. However, norms are subject to change, as Danescu-Niculescu-Mizil et al. (2013b) have shown, and

[2] Note that the latter two show that human speakers depend on context as well, though.

[3] https://www.theguardian.com/world/2016/mar/29/microsoft-tay-tweets-antisemitic-racism

[4] https://www.theguardian.com/world/2021/jan/14/time-to-properly-socialise-hate-speech-ai-chatbot-pulled-from-facebook

can affect standing and integration of members.

2.6 Culture and Ideology

Language and culture are intertwined. Language reflects the society, ideology, cultural identity, and customs of communicators, as well as their values. It is therefore intertwined with social norms (Section 2.5). For example, in Japanese (Gao, 2005), the expression of hierarchy necessitates more fine-grained politeness and formality levels than in Western cultures. The terms of address also vary in terms of social and age differences, i.e., inferior members address superior ones with a relationship term instead of using personal names (see also Section 2.3). In many Asian cultures, family terms like "*uncle*" or "*big sister*" are used as honorifics. While it is common amongst native speakers of North American English to use "*please*" in requests even to close friends, such an act would be considered awkward, if not rude, in Arabic-speaking cultures (Kádár and Mills, 2011; Madaan et al., 2020).

Cultural norms can impose a hierarchy on Gricean maxims. For example, whether it is better to give made-up directions (which violates the maxim of relevance) instead of not saying anything (adhering to the maxim of quality) if you do not know the right answer.

Context and social and cultural norms can combine in unexpected ways, such as in the case of Korean Airline co-pilots not correcting pilot mistakes (a social and cultural taboo in ordinary contexts), which resulted in a series of accidents. Differing perceptions of the context, respect for seniority and age, and a hierarchical communication style can lead to one-way communication, in these cases resulting in the deaths of hundreds.[5] The solution here was to change the context by making the working language English, which in turn removed associated social and cultural norms around hierarchical communication (Gladwell, 2008).

Applications Culture and ideology are probably the most complicated language constructs. Despite their substantial influence on communication interpretation and language understanding, most NLP models, like text generation or translation, have not included politeness or other similar subtle cultural signatures. A growing body of research has paid attention to the biases and cultural stereotypes encoded and amplified by current NLP models, e.g., inappropriate occupation predictions by large pretrained language models like "*the black woman who worked as a babysitter*" (Sheng et al., 2019). These findings call for work to look at the ideology, beliefs, and culture behind language content to mitigate biases and social stereotypes beyond data-level manifestations. The fact that embeddings reflect these stereotypes, cultural beliefs, and ideologies make them also an ideal diagnostic tool for social science scholars (Garg et al., 2018; Kozlowski et al., 2018). However, it also creates fundamental biases that cannot easily be mitigated (Gonen and Goldberg, 2019), which poses severe problems for their use in predictive models. Adding cultural awareness can also help counteract the overexposure (Hovy and Spruit, 2016) to the English language (Joshi et al., 2020)[6] and Anglo-Western culture.

2.7 Communicative Goal

Finally, communicative goals cover what people want to achieve with their language use, e.g., information, decision making, social chitchat, negotiation, etc. SFL represents this factor as multiple metafunctions of language. Two metafunctions are of particular relevance here: the *interpersonal* metafunction, whereby language enables us to enact social relationships, to cooperate, form bonds, negotiate, ask for things, and instruct; and the *ideational* metafunction, whereby language enables us to talk *about* inner and outer experiences, people and things, or circumstances in which events occur. Goals introduce an essential layer on top of content, and a good understanding of them can reveal the intent and implication behind the text structure. All of the Gricean maxims are used (or deliberately flaunted) in the service of achieving these goals. For example, when trying to convince someone to join us in a project, we might adhere to the maxims of relevance and concisely lay out the reasons we need them to join. However, to make it more likely that they agree, we might choose to exaggerate the expected payoff and to leave out some of the difficulties involved, which violates the maxims of quality and quantity, respectively.

Applications Communicative goals shape how speakers arrange their words and styles. For in-

[5] https://www.cnbc.com/id/100869966

[6] https://thegradient.pub/the-benderrule-on-naming-the-languages-we-study-and-why-it-matters/

stance, text that aims to convince others often uses various persuasion strategies (Yang et al., 2019a; Chen and Yang, 2021), argumentation techniques (Stab and Gurevych, 2014), rhetorical structures (Rapp, 2011), and the exchange of social support (Wang and Jurgens, 2018; Yang et al., 2019b). Messages trying to entertain audiences need to be structured in ways that can trigger humor (Yang et al., 2015). People might use informal language or text with a high level of intimacy to indicate close relations (Pei and Jurgens, 2020) or reduce social distance between speakers and receivers (Bernstein, 1960; Keshavarz, 2001).

Therefore, it is essential for NLP systems like text generation models to be aware of communicative goals in order to arrange word choice, and styles to form a grammatically responsible and coherent text. Ongoing research has shown that style can be controlled independently of content(Prabhumoye et al., 2018; John et al., 2019). Some of the early work on NLP (Hovy, 1987) explicitly considered communicative goals in sentence generation, albeit modeled explicitly. More recently, Sap et al. (2020) modeled speaker intent to infer intent and resolve conversational implicature.

3 Outlook and Challenges

Social Factors in Different NLP Tasks When and how, though, should we consider these various social factors for an NLP application? NLP practitioners should feel free to use our social factor taxonomy as a guide to examine what social factors should be used, and whether integrating each confers additional benefits (e.g., better design, performance, user experience, or cultural fit) for their use cases. Different NLP tasks will likely benefit differently from our social factor taxonomy.

There is some evidence that the earlier factors (such as speaker and receiver characteristics) can be applied to most tasks, as they are fundamental aspects of language. Social relations and context are likely to apply more to dialogue and text generation tasks than to, say, sentiment analysis. Lastly, "high-level" factors such as social norms and culture and ideology likely require more research to inform individual applications, but are likely to shape our community approaches. We would be well-advised to incorporate the findings of fields that have studied these issues for longer, such as philosophy, sociology, or sociolinguistics. As NLP tasks and algorithms are being now applied to different aspects of everyday interaction and around the world, how we will equip NLP models with a grounding in social factors becomes extremely important, especially these two dimensions. Detailed modeling of these social factors is essential if NLP systems are to have any impact. It can also help avoid hegemonic approaches from assuming all conversations follow Western norms, culture, and ideology.

Real-world interaction involves more than the exchange of information or decision making via language; it involves a wide range of aspects related to social factors and interpersonal relations, reflected in rich modalities such as voice or facial expression. Though this work's focus is on the language side, we argue that the introduced taxonomy can be beneficial in broader scenarios for next-level multi-modal models.

Data, Ethics, and Privacy Our work here is related to some of the recent work on bias in NLP (Hovy and Spruit, 2016; Shah et al., 2020). On the one hand, the cooperative principle can be seen as a possible positive bias: a pre-existing expectation of how we interact, the violation of which signals an alternative approach. So far, models do not integrate this positive bias. On the other hand, work on speaker and receiver characteristics is affected by the models' predictive biases: exaggerating or overestimating one particular group's attributes can skew the results, for example, in the case of machine-translated texts sounding older and more male (Hovy et al., 2020). Recently, Blodgett et al. (2020) have discussed the role of "bias" conceptions, which serves as a meta-discussion of the conceptualization of social norms.

Integrating social factors into NLP poses a double challenge: on the one hand, it requires additional data to model those social factors. We need representative annotation samples for, e.g., the demographics and network information of speaker, receiver, and social relations, which requires us to collect and document our annotations (Bender and Friedman, 2018). Social media already contains some information from personal or socially grounded conversations, but other domains might suffer from data sparsity for these factors, and require advances in unsupervised learning or few-shot learning techniques.

On the other hand, collecting all this information raises questions about privacy, data protection, and ethics. Some data we need to collect to work with social factors is personal or protected data, which comes with risks for de-anonymization and privacy leaks. Collecting sensitive data (i.e., membership in a protected category) requires the participants' approval and procedures to ensure that this information cannot be connected to them individually. These considerations also pose a challenge to data sharing; even if properly anonymized, data can contain clues as to participants' identity(Eckert and Dewes, 2017). We will need to strengthen ethical considerations for this emerging direction to guide practice in the field and ensure our models are used in beneficial ways.

Evaluation and Metrics A central question in these efforts is *How do we evaluate whether NLP models have learned the social factors of language, beyond performance improvements?* Current models optimize performance metrics, but these metrics might fail to capture the nuances of NLP systems' understanding when considering social content. Thus, better metrics are needed to measure and visualize such additional benefits introduced by modeling language's social factors. These metrics will become essential to diagnose failure. Failed or improper incorporation of social factors could lead to awkward social consequences. E.g., a system misjudging its social relation to the speaker and being a bit too "chummy", or a conversational agent disrespecting social norms of turn taking and formality. To some extent, such problems might be unavoidable: interacting through language is always a trial-and-error process, even for humans. However, such "errors" become extremely important in high-stakes scenarios, such as inappropriate responses from conversational agents in mental health counseling applications. We need metrics to capture this failure and mechanisms to explain the decision-making process behind socially aware NLP models.

Multi-modal Social Interaction Real-world interaction involves more than the exchange of information or decision making via language; it involves a wide range of aspects related to social factors and interpersonal relations, reflected in rich modalities (Simmons et al., 2011) such as images, voice or facial expression. Though this work's focus is on the language side, we argue that the introduced taxonomy can be beneficial in broader scenarios for the next level multi-modal models.

4 Conclusion

In this work, we have argued that there are seven social factors of language that impact NLP applications: speaker, receiver characteristics, social relations, context, social norms, culture and ideology, and communicative goals. At present, NLP models often ignore these factors. We have shown that this ignorance limits the kinds of applications we can tackle. It can also can introduce mistakes, ranging from the hilarious to the severe. However, several extant approaches incorporate these social factors, all of them showing substantial improvements in a wide range of applications. By systematically addressing the social aspects of language as a field, we will improve the performances of existing NLP systems, open up new applications, and increase fairness and usability for all users.

Acknowledgements

This project has partially received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (No. 949944, INTEGRATOR). DY is supported in part by grants from Google and Salesforce. DH is the scientific director of the Data and Marketing Insights Unit at the Bocconi Institute for Data Science and Analysis. We would like to thank Maxwell Forbes, Christoph Purschke, and Maarten Sap for comments on the drafts, as well as the three reviewers who suggested various additions.

References

Gavin Abercrombie and Dirk Hovy. 2016. Putting sarcasm detection into context: The effects of class imbalance and manual labelling on supervised machine classification of Twitter conversations. In *Proceedings of the ACL 2016 Student Research Workshop*, pages 107–113, Berlin, Germany. Association for Computational Linguistics.

Nicolas Anastassacos, Stephen Hailes, and Mirco Musolesi. 2020. Partner selection for the emergence of cooperation in multi-agent systems using reinforcement learning. In *AAAI*, pages 7047–7054.

Rafael E Banchs and Haizhou Li. 2012. Iris: a chat-oriented dialogue system based on the vector space model. In *Proceedings of the ACL 2012 System Demonstrations*, pages 37–42. Association for Computational Linguistics.

Justin Basilico and Thomas Hofmann. 2004. Unifying collaborative and content-based filtering. In *Proceedings of the twenty-first international conference on Machine learning*, page 9.

Emily M Bender and Batya Friedman. 2018. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Adrian Benton, Margaret Mitchell, and Dirk Hovy. 2017. Multitask learning for mental health conditions with limited social media data. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 152–162, Valencia, Spain. Association for Computational Linguistics.

Basil Bernstein. 1960. Language and social class. *The British journal of sociology*, 11(3):271–276.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. 2020. Experience grounds language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8718–8735, Online. Association for Computational Linguistics.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic dialectal variation in social media: A case study of African-American English. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1119–1130, Austin, Texas. Association for Computational Linguistics.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Ozan Caglayan, Pranava Swaroop Madhyastha, Lucia Specia, and Loïc Barrault. 2019. Probing the need for visual context in multimodal machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4159–4170.

Amanda Cercas Curry and Verena Rieser. 2018. #MeToo Alexa: How conversational systems respond to sexual harassment. In *Proceedings of the Second ACL Workshop on Ethics in Natural Language Processing*, pages 7–14, New Orleans, Louisiana, USA. Association for Computational Linguistics.

Amanda Cercas Curry, Judy Robertson, and Verena Rieser. 2020. Conversational assistants and gender stereotypes: Public perceptions and desiderata for voice personas. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 72–78, Barcelona, Spain (Online). Association for Computational Linguistics.

Eshwar Chandrasekharan, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Cliff Lampe, Jacob Eisenstein, and Eric Gilbert. 2018. The internet's hidden rules: An empirical study of reddit norm violations at micro, meso, and macro scales. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–25.

Jiaao Chen and Diyi Yang. 2021. Weakly-supervised hierarchical models for predicting persuasive strategies in good-faith textual requests. *AAAI*.

Herbert H. Clark and Michael F. Schober. 1992. Asking questions and influencing answers. *Questions about Questions: Inquiries into the Cognitive Bases of Surveys*, pages 15–48.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of machine learning research*, 12(ARTICLE):2493–2537.

Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013a. A computational approach to politeness with application to social factors. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 250–259, Sofia, Bulgaria. Association for Computational Linguistics.

Cristian Danescu-Niculescu-Mizil, Robert West, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013b. No country for old members: User lifecycle and linguistic change in online communities. In *Proceedings of the 22nd international conference on World Wide Web*, pages 307–318.

Marco Del Tredici, Diego Marcheggiani, Sabine Schulte im Walde, and Raquel Fernández. 2019. You shall know a user by the company it keeps: Dynamic representations for social media users in nlp. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4701–4711.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Emily Dinan, Angela Fan, Ledell Wu, Jason Weston, Douwe Kiela, and Adina Williams. 2020. Multi-dimensional gender bias classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 314–331, Online. Association for Computational Linguistics.

Rachel Dorn. 2019. Dialect-specific models for automatic speech recognition of African American Vernacular English. In *Proceedings of the Student Research Workshop Associated with RANLP 2019*, pages 16–20, Varna, Bulgaria. INCOMA Ltd.

Penelope Eckert. 2012. Three waves of variation study: The emergence of meaning in the study of sociolinguistic variation. *Annual review of Anthropology*, 41:87–100.

Svea Eckert and Andreas Dewes. 2017. Dark data. *Presentation at DEFCON*, 25.

Marcello Federico. 1996. Bayesian estimation methods for n-gram language model adaptation. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, volume 1, pages 240–243. IEEE.

Ernst Fehr and Urs Fischbacher. 2004. Social norms and human cooperation. *Trends in cognitive sciences*, 8(4):185–190.

Charles J Fillmore. 1982. Frame semantics. In *Linguistics in the morning calm*, pages 111–137.

Lucie Flek. 2020. Returning the N to NLP: Towards contextually personalized classification models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7828–7838, Online. Association for Computational Linguistics.

Maxwell Forbes, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 2020. Social chemistry 101: Learning to reason about social and moral norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 653–670, Online. Association for Computational Linguistics.

Tommaso Fornaciari and Dirk Hovy. 2019. Dense node representation for geolocation. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 224–230, Hong Kong, China. Association for Computational Linguistics.

Liye Fu, Susan Fussell, and Cristian Danescu-Niculescu-Mizil. 2020. Facilitating the communication of politeness through fine-grained paraphrasing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5127–5140, Online. Association for Computational Linguistics.

Fengping Gao. 2005. Japanese: A heavily culture-laden language. *Journal of Intercultural Communication*, 10:1404–1634.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Malcolm Gladwell. 2008. *Outliers: The story of success*. Little, Brown.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 609–614.

Herbert P Grice. 1975. Logic and conversation. In *Speech acts*, pages 41–58. Brill.

Michael Alexander Kirkwood Halliday and Christian MIM Matthiessen. 2013. *Halliday's introduction to functional grammar*. Routledge.

Dirk Hovy. 2015. Demographic factors improve classification performance. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 752–762, Beijing, China. Association for Computational Linguistics.

Dirk Hovy. 2016. The Enemy in Your Own Camp: How Well Can We Detect Statistically-Generated Fake Reviews–An Adversarial Study. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Dirk Hovy. 2018. The social and the neural network: How to make natural language processing about people again. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pages 42–49, New Orleans, Louisiana, USA. Association for Computational Linguistics.

Dirk Hovy, Federico Bianchi, and Tommaso Fornaciari. 2020. "you sound just like your father" commercial machine translation systems include stylistic biases. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1686–1690.

Dirk Hovy and Christoph Purschke. 2018. Capturing regional variation with distributed place representations and geographic retrofitting. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4383–4394, Brussels, Belgium. Association for Computational Linguistics.

Dirk Hovy and Shannon L Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 591–598.

Eduard Hovy. 1987. Generating natural language under pragmatic constraints. *Journal of Pragmatics*, 11(6):689–719.

Dell Hymes. 1972. On communicative competence. *sociolinguistics*, 269293:269–293.

Mohit Iyyer, Anupam Guha, Snigdha Chaturvedi, Jordan Boyd-Graber, and Hal Daumé III. 2016. Feuding families and former friends: Unsupervised learning for dynamic fictional relationships. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1534–1544.

Vineet John, Lili Mou, Hareesh Bahuleyan, and Olga Vechtomova. 2019. Disentangled representation learning for non-parallel text style transfer. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 424–434, Florence, Italy. Association for Computational Linguistics.

Ruth Jones and Ann Irvine. 2013. The (un)faithful machine translator. In *Proceedings of the 7th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 96–101, Sofia, Bulgaria. Association for Computational Linguistics.

Anna Jørgensen, Dirk Hovy, and Anders Søgaard. 2016. Learning a POS tagger for AAVE-like language. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1115–1120, San Diego, California. Association for Computational Linguistics.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Prathyusha Jwalapuram. 2017. Evaluating dialogs based on Grice's maxims. In *Proceedings of the Student Research Workshop Associated with RANLP 2017*, pages 17–24, Varna. INCOMA Ltd.

Dániel Z Kádár and Sara Mills. 2011. *Politeness in East Asia*. Cambridge University Press.

Mohammad Hossein Keshavarz. 2001. The role of social context, intimacy, and distance in the choice of forms of address. *International journal of the sociology of language*, 2001(148):5–18.

Sung-wan Kim and HyoJung Lee. 2017. A study on machine translation outputs: Korean to english translation of embedded sentences. 22(4):123–147.

Austin C Kozlowski, Matt Taddy, and James A Evans. 2018. The geometry of culture: Analyzing meaning through word embeddings. *arXiv preprint arXiv:1803.09288*.

Vinodh Krishnan and Jacob Eisenstein. 2015. "You're mr. Lebowski, I'm the Dude": Inducing address term formality in signed social networks. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1616–1626, Denver, Colorado. Association for Computational Linguistics.

William Labov. 1972. *Sociolinguistic patterns*. University of Pennsylvania Press.

Angeliki Lazaridou, Nghia The Pham, and Marco Baroni. 2015. Combining language and vision with a multimodal skip-gram model. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 153–163, Denver, Colorado. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 994–1003.

Yitong Li, Timothy Baldwin, and Trevor Cohn. 2018. Towards robust and privacy-preserving text representations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 25–30, Melbourne, Australia. Association for Computational Linguistics.

Eden Litt. 2012. Knock, knock. who's there? the imagined audience. *Journal of broadcasting & electronic media*, 56(3):330–345.

Veronica Lynn, Youngseo Son, Vivek Kulkarni, Niranjan Balasubramanian, and H. Andrew Schwartz. 2017. Human centered NLP with user-factor adaptation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1146–1155, Copenhagen, Denmark. Association for Computational Linguistics.

Aman Madaan, Amrith Setlur, Tanmay Parekh, Barnabas Poczos, Graham Neubig, Yiming Yang, Ruslan Salakhutdinov, Alan W Black, and Shrimai Prabhumoye. 2020. Politeness transfer: A tag and generate approach. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1869–1881, Online. Association for Computational Linguistics.

Christopher D Manning. 2015. Computational linguistics and deep learning. *Computational Linguistics*, 41(4):701–707.

Evgeny Matusov. 2019. The challenges of using neural machine translation for literature. In *Proceedings of the Qualities of Literary Machine Translation*, pages 10–19, Dublin, Ireland. European Association for Machine Translation.

Shachar Mirkin and Jean-Luc Meunier. 2015. Personalized machine translation: Predicting translational preferences. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2019–2025.

Dong Nguyen, A. Seza Doğruöz, Carolyn P. Rosé, and Franciska de Jong. 2016. Survey: Computational sociolinguistics: A Survey. *Computational Linguistics*, 42(3):537–593.

Debora Nozza, Federico Bianchi, and Dirk Hovy. 2021. HONEST: Measuring hurtful sentence completion in language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics*, Minneapolis, Minnesota. Association for Computational Linguistics.

Jiaxin Pei and David Jurgens. 2020. Quantifying intimacy in language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5307–5326.

Vinodkumar Prabhakaran, Owen Rambow, and Mona Diab. 2012. Predicting overt display of power in written dialogs. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 518–522. Association for Computational Linguistics.

Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W Black. 2018. Style transfer through back-translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 866–876, Melbourne, Australia. Association for Computational Linguistics.

Mohammed R. H. Qwaider, Abed Alhakim Freihat, and Fausto Giunchiglia. 2017. TrentoTeam at SemEval-2017 task 3: An application of Grice maxims in ranking community question answers. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 271–274, Vancouver, Canada. Association for Computational Linguistics.

Bahar Radfar, Karthik Shivaram, and Aron Culotta. 2020. Characterizing variation in toxic language by social context. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 959–963.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.

Afshin Rahimi, Trevor Cohn, and Timothy Baldwin. 2018. Semi-supervised user geolocation via graph convolutional networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2009–2019, Melbourne, Australia. Association for Computational Linguistics.

Christof Rapp. 2011. Aristotle's rhetoric. *Stanford Encyclopedia of Philosophy*.

Farzana Rashid and Eduardo Blanco. 2017. Dimensions of interpersonal relationships: Corpus and experiments. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2307–2316.

Farzana Rashid, Tommaso Fornaciari, Dirk Hovy, Eduardo Blanco, and Fernando Vega-Redondo. 2020. Helpful or hierarchical? predicting the communicative strategies of chat participants, and their impact on success. In *Findings of the Association for Computational Linguistics: Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5248–5264, Online. Association for Computational Linguistics.

Hannah Rashkin, Sameer Singh, and Yejin Choi. 2016. Connotation frames: A data-driven investigation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 311–321, Berlin, Germany. Association for Computational Linguistics.

Alan Ritter, Colin Cherry, and William B. Dolan. 2011. Data-driven response generation in social media. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 583–593, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019a. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th*

Annual Meeting of the Association for Computational Linguistics, pages 5477–5490, Online. Association for Computational Linguistics.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019b. Social iqa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4453–4463.

Roger C Schank and Robert P Abelson. 1975. Scripts, plans, and knowledge. In *IJCAI*, volume 75, pages 151–157.

Iulian V Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Deven Santosh Shah, H. Andrew Schwartz, and Dirk Hovy. 2020. Predictive biases in natural language processing models: A conceptual framework and overview. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5248–5264, Online. Association for Computational Linguistics.

Xuehua Shen, Bin Tan, and ChengXiang Zhai. 2005. Implicit user modeling for personalized search. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 824–831.

Yelong Shen, Xiaodong He, Jianfeng Gao, Li Deng, and Grégoire Mesnil. 2014. Learning semantic representations using convolutional neural networks for web search. In *Proceedings of the 23rd international conference on world wide web*, pages 373–374.

Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3398–3403.

Michael Silverstein. 2003. Indexical order and the dialectics of sociolinguistic life. *Language & communication*, 23(3-4):193–229.

Matthew Simmons, Lada Adamic, and Eytan Adar. 2011. Memes online: Extracted, subtracted, injected, and recollected. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 5.

Christian Stab and Iryna Gurevych. 2014. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 46–56.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy. Association for Computational Linguistics.

Peter Trudgill. 2000. *Sociolinguistics: An introduction to language and society*. Penguin UK.

Amos Tversky and Daniel Kahneman. 1973. Availability: A heuristic for judging frequency and probability. *Cognitive psychology*, 5(2):207–232.

Eva Vanmassenhove, Dimitar Shterionov, and Andy Way. 2019. Lost in translation: Loss and decay of linguistic richness in machine translation. In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*, pages 222–232, Dublin, Ireland. European Association for Machine Translation.

Svitlana Volkova, Theresa Wilson, and David Yarowsky. 2013. Exploring demographic language variations to improve multilingual sentiment analysis in social media. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1815–1827.

Byron C. Wallace, Do Kook Choe, Laura Kertz, and Eugene Charniak. 2014. Humans require context to infer ironic intent (so computers probably do, too). In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 512–516, Baltimore, Maryland. Association for Computational Linguistics.

Zijian Wang and David Jurgens. 2018. It's going to be okay: Measuring access to support in online communities. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 33–45.

William Warner and Julia Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the Second Workshop on Language in Social Media*, pages 19–26, Montréal, Canada. Association for Computational Linguistics.

Bolin Wei, Shuai Lu, Lili Mou, Hao Zhou, Pascal Poupart, Ge Li, and Zhi Jin. 2017. Why do neural dialog systems generate short and meaningless replies? A comparison between dialog and translation. *CoRR*, abs/1712.02250.

Ludwig Wittgenstein. 2010. *Philosophical investigations*. John Wiley & Sons.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Yuwei Wu, Xuezhe Ma, and Diyi Yang. 2021. Personalized Response Generation via Generative Split Memory Network. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Diyi Yang. 2019. *Computational Social Roles*. Ph.D. thesis, Ph. D. thesis, Carnegie Mellon University.

Diyi Yang, Jiaao Chen, Zichao Yang, Dan Jurafsky, and Eduard Hovy. 2019a. Let's make your request more persuasive: Modeling persuasive strategies via semi-supervised neural nets on crowdfunding platforms. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3620–3630, Minneapolis, Minnesota. Association for Computational Linguistics.

Diyi Yang, Robert E Kraut, Tenbroeck Smith, Elijah Mayfield, and Dan Jurafsky. 2019b. Seekers, providers, welcomers, and storytellers: Modeling social roles in online health communities. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–14.

Diyi Yang, Alon Lavie, Chris Dyer, and Eduard Hovy. 2015. Humor recognition and humor anchor extraction. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2367–2376, Lisbon, Portugal. Association for Computational Linguistics.

Yi Yang and Jacob Eisenstein. 2017. Overcoming language variation in sentiment analysis with social attention. *Transactions of the Association for Computational Linguistics*, 5:295–307.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489, San Diego, California. Association for Computational Linguistics.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.