Article

Towards an Android Linguistics: Pragmatics, Reflection, and Creativity in Machine Language

Evan Donahue

Tokyo College, University of Tokyo; evan.donahue@tc.u-tokyo.ac.jp

Abstract: Contemporary natural language processing (NLP) emphasizes comparing machine language performances to standards defined by static corpora of human text. However, despite some successes, current models remain weak in areas such as pragmatics. Using scholarship on neosentience as a point of departure, this essay proposes an alternative view of machine language that emphasizes generativity rather than stasis, and draws on historical work on computational reflection in artificial intelligence to sketch an alternative architecture for conversational systems. It concludes by proposing an "android linguistics" that takes human-machine linguistic communication as its object of study.

Keywords: artificial intelligence; natural language processing; cybernetics; neosentience; pragmatics; speech acts; reflection; self-reference

Citation: Donahue, E. Towards an Android Linguistics: Pragmatics, Reflection, and Creativity in Machine Language. *Proceedings* **2021**, *1*, 0. https://doi.org/

Published:

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Introduction

Is language a noun or a verb? Contemporary NLP views language as a static target defined by human performances and against which machine performances must be measured. Such work views language implicitly as a static encyclopedia within which can be found the appropriate response to any future conversational context [1]. The cybernetically-inflected view of language offered by neosentience scholarship and recombinant poetics offers a useful contrast in viewing language use as an act of creation, and offers a different way to think about designing conversational machines [2].

Expanding on the neosentient view of language as an act of creation, I suggest that machine language researchers' perennial difficulties with pragmatics—with accounting for the influence of context on interpretation—cannot be solved with scale, but instead require a different, self-reflexive architecture. Pragmatics, I contend, is inseparable from self-reflection, and in this essay I motivate the need for such an architecture, suggest some preliminary requirements for its design, and call for deeper consideration of language as a phenomenon not limited to human speakers, with implications for how we decide what counts as linguistic performance by a machine.

Pragmatics as Self-Reflection

AI researcher Douglas Lenat, although supportive of scaling up language systems, tempers his support with a note of caution. In observing how the slightest change in the placement of a comma can radically alter the interpretation of a sentence by inviting in a host of assumptions about the context in and the purpose for which it was written, he writes despairingly that, "there's always this annoying residue of pragmatics, which ends up being the lower 99% of the iceberg. . . lurking in the empty spaces around the letters, words, and sentences" [3, 2]. Like dark matter, pragmatics constitutes for Lenat the vast and unseen majority of the reality of language.

That such a reality should so trouble Lenat speaks to the seeming hopelessness of fitting the inexhaustible totality of language into a finite computer system. Through the lens of recombinant poetics, however, the infinite creativity of language is precisely what makes conversation possible. Shifting the focus of machine language research from attempting to build machines that know in advance what words mean to negotiating that meaning within the context of their creation, however, requires a shift in our understanding of language itself in the machinic context.

If generativity rather than stasis is taken to be the nature of language, then language systems must be designed less to know vast quantities of language than to attend reflexively to the context of its creation. As Seaman [4] argues, agents in conversation generate meaning not so much by dredging it out of a database fully formed but rather by attending to their own and others' efforts to bring it into being with the words, gestures, and objects available to them. By existing and conversing, a conversational agent creates new utterances and new meanings that can never be captured by the dataset of other speakers on which it was trained. The one thing the dataset cannot contain, no matter how vast, is its own model, and yet it is only by understanding itself and the context of its own speech as implicated in the process of meaning-making that the agent can make sense of the new meanings made.

Self-reflexivity has a long history in studies of machine intelligence. The idea that a computer system could be taught to rewrite its own programming, and thereby exceed the limitations imposed by human-engineered systems, has attained at times an almost mystical quality in artificial intelligence. Seaman [5] focuses on a more contained form of self-reference in language through a discussion of Givón's work on pragmatics. As Givón [6] argues, it is language's ability to refer to itself in discourse that underlies much of its pragmatic functioning. Correspondingly, I contend that the central architectural principle that must underwrite any attempt at building a pragmatically competent conversational system is that the objects of discourse itself—the linguistic inputs to and outputs from the system—must themselves be first-class objects within that discourse. It must be possible, using computational reflection, to move fluidly between considering an utterance as a meaning-bearing element of conversation and viewing it as an object about which other meaning-bearing utterances can be made. From this principle, much of the rest of what I believe is important for pragmatic computer systems will follow.

Pragmatics in AI

AI inherited from philosophy a Russellian epistemology in which sentences map neatly onto reality. Words refer to objects or actions, statements are true or false depending on whether the words that comprise them correspond to this reality, and questions can be answered correctly or incorrectly depending on whether their correspondences with that reality align with those of their answers. Consequently, self-reflexive statements such as "this statement is false" that frustrate efforts to parse them as either true or false have long haunted the field's efforts to conceptualize meaning in machine language.

A pragmatic explanation for this phenomenon is that the human, unlike most AI systems, escapes to another level of interpretation. While perfectly capable of playing the language game of true and false, albeit without the precision of the computer, the human is also capable of changing perspectives and recognizing such a self referential statement as an utterance made by a logician attempting to prove a point about a particular language game, as in this very essay, even though the logician making the point appears nowhere in the statement. AI researchers have long intuited this change of levels, as when Marvin Minsky observes that the usual reaction to such "liar's paradoxes" is not to spin endlessly trying to solve them, but rather to laugh and reject the language game of true and false altogether [7, 139]. In an instant, the paradoxical statement reduced to a sequence of sounds that have no meaning and present no paradox.

Richard Weyhrauch, whose work with Carolyn Talcott and others on reflective computer architectures was animated by the parallels between self reference and conversation, remarks that much of what we would like to talk to machines about is language itself [8, 155]. Although he does not invoke pragmatics explicitly, his illustrations of the contextual nature of meaning foreground its centrality in thought on machine communication. As he observes, the truth value of even the most seemingly inescapable universal proposition, such as "2+2=4," becomes suddenly suspect when growled angrily across a bar table by the head of a biker gang [9, 14]. Even the staunchest defender of universal meanings would likely have to think twice about whether they were in fact caught up in some larger language game that subverted the apparent meaning of the statement.

Weyhrauch's example highlights the key weakness of any attempt to design or evaluate machine language by sorting the world into clear categories against which to test the performance of AI systems. Even the most intuitively unambiguous utterances can always have their meanings utterly displaced in the context of the right language game. As in a spy novel, it is always possible for a seemingly innocuous utterance to signify the transmission of a secret code. Whether such language games represent exceptional circumstances that can be safely ignored until the basics of machine language have been figured out or whether they point to more fundamental mechanisms the omission of which will doom the whole project is the question that must be addressed and the point on which poetic and encyclopedic theories of machine language most differ.

This difference becomes apparent when considering even the simplest of sentences. A recent project by researchers at Facebook offers one perspective on the nature of machine language by attempting to enumerate a set of tests of fundamental reasoning abilities a machine must possess to answer simple questions [10]. These tests include an understanding of true and false, of elements and sets, of logical conjunction and disjunction, of negation, and of numbers. Taken together, these tests assess an artificial agent's ability to emulate a formal reasoning system. While not necessarily an unreasonable capacity to ask of an artificial agent, from the point of view of pragmatics it omits a more fundamental level of analysis. Namely, it rests on an unproblematic mapping from sentences of English to sentences of logic. The sentences themselves are not objects of discourse within the proposed test environment to be reasoned about.

The importance of being able to speak at this meta-level of discourse becomes evident in consideration of Searle's famous question, "can you pass the salt?" [11]. A system trained to respond to questions may go wrong if it assumes that the question is a request for information about the machines capabilities. The purpose of this question is of course to illustrate that what appears to be a question about passing salt may in fact be a request to effect the passing. Then again, in another context the question may not even be asked in good faith but rather as part of a test of the system's understanding of pragmatics, perhaps designed in response to an essay such as this one, in which case the "correct" response would depend on the level at which one interprets the scene.

Any test of an AI's linguistic competence necessarily rests on a set of assumptions on the part of the researcher about the context in which the language is to be interpreted and what correctness or incorrectness looks like in such a context. From the system's point of view, however, it is never told it is being evaluated. The parameters of the evaluation are never explained to the system in language, in part because in most cases it lacks the representational machinery to even recognize the discursive elements of a language test as entities with which it shares a reality. It is in the position of the chess playing automaton that does not know it is playing chess; it may play a fair game, but a human operator must carry it in, set it up next to the board, and face it the right direction. As long as this remains the default experimental paradigm in AI, it is likely that special purpose systems that exhibit virtuoso linguistic performances without therefore becoming communicatively competent will continue to be the norm.

Searle's question underscores that the sign is not just arbitrary in the Saussurean sense that there is no necessary relationship between the form of the signifier and the signified it has historically come to represent, but that it is radically arbitrary. No matter the history of the signifier or the conventions of its interpretation in other contexts, it is always possible to subvert that history with the appropriate language game in the appropriate context. The neosentient view suggests that this subversion happens rapidly and continuously as part of the creative flux of language in practice. Starting from this intuition, any attempt to learn the correspondence between the form of the utterance and its meaning is to build castles on sand. What is needed is an approach that underlies and precedes the performances the Facebook researchers seek to measure that stages the act of interpretation as prior to the consideration of form.

Towards an Android Linguistics

Contemporary NLP depends heavily on quantifying the correctness of language. However, such quantification is a compromise with which perhaps no one in the field has ever been truly happy [12]. Moreover, the field's history offers a wealth of examples of alternative conceptions of the work of studying machine language that may offer inspiration. In particular, with respect to the question of pragmatics, a body of work responding to the speech act theory of Searle, Austin, and others that emerged in the late 1970s offers a distinctive approach to conceptualizing machine language [13–16].

While a full review of this literature is beyond the scope of this essay, one key insight that offers concrete guidance on the design and evaluation of contemporary systems is that form must be held entirely apart from meaning. "Can you pass the salt" should not immediately resolve into either a request for information or a request for action. Rather, the words must be evaluated based on what is known about the speaker and the environment for what they might indicate about the beliefs, intentions, and desires of that speaker. Once it is determined that the speaker desires the salt and believes the system can obtain it for them, then the system can exercise its agency by passing the salt or witholding it, by speaking or remaining silent. This action is undertaken not on the basis of the force of the utterance itself but on that of what the utterance has revealed about the interiority of its utterer, and that revelation is a product not only of interpretation but of conversational interaction—of the poiesis of meaning—enabled by the explicit self-representation of the field of discourse.

Several important conversational behaviors are enabled by representing discursive objects explicitly alongside other domain objects in a reflective manner. These might stand alongside the behaviors outlined by the Facebook researchers as heuristics with which to probe whether a conversational system is representationally sufficient to capture even in principle important pragmatic dimensions of communication.

The most basic requirement for a pragmatic conversational system is the ability to refer to the words of the conversation itself. Many current systems, if they had never encountered the word "salt," would simply fail to process Searle's question, rather than being able to formulate a question of which the word itself was the subject. Even in asking for clarification of how a known word or construction is being used creatively in a new context requires the ability to refer to the word in question.

The second property is the ability to refer to the system's own interpretations of prior utterances as first-class discursive objects. Inevitably, in conversation, misunderstandings will arise. Repairing them hinges in part on the ability to discuss what was previously understood in order to create a new context for further conversation. Explaining to the system that Searle's question was intended as a request for salt rather than for information hinges on the ability to ground references to its erroneous interpretation that the speaker wanted only a verbal response, even if such an interpretation is only a discursive fiction rather than a technical reality in the underlying system.

Repairing the conversation, in turn, is not a simple matter of correcting a previous misunderstanding to what it originally should have been, but rather of determining how that meaning has been changed in light of statements made and actions taken on the basis of the misunderstanding. In order to ask about how to proceed, and whether the speaker still wants the salt, the system must be able to discuss its plans and adjust them based on the conversation that follows. Discussion of such plans in turn requires the ability to refer to the future worlds that such plans might bring about. Although the mental or physical reality of plans and possible worlds has been a longstanding point of debate within AI, their reality as objects of the discursive universe requires no underpinning outside of language to validate their utility.

Finally, the ability to project hypothetical possible worlds invites consideration of the ability to project fictional ones—worlds that do not exist and that could not in principle exist or that may not even make complete sense. If an AI system were to read a work of fiction, it should be able to do so without either confusing it with the world beyond the fiction or keeping the two so wholly separate that the linguistic and cultural materials with which fictions are constructed become unintelligible. Moreover, as work on narrative theory and story worlds seems to hint, it may be worth viewing reality itself as a collection of fictions we tell and retell, inventing in the process that which can never be captured by a single totalizing view of language as such [17,18].

Conclusion

To treat language as a static whole rather than a dynamic process in which the researchers themselves are implicated is, to paraphrase Givón, to rescue the study of machine language by abandoning its purpose [6, 4]. Indeed, any artificially intelligent system not intelligent enough to know it is being tested is unworthy of the name. Centering context in communication promises more conversationally capable machines, and this essay has offered as a starting point the narrow technical requirement that objects of discourse should have first-class representations in the system. More speculatively, because consideration of context calls attention to speakers and their standpoints, it is perhaps worth contemplating whether interrogating how it normalizes certain language as "correct" might force the field to confront the assumptions about race, gender, and disability inscribed on its datasets and artifacts that scholars have documented for decades [19,20]. This attention to the language of machines and its place in broader human language communities as a basis for the design of socio-technical systems is what I refer to as an android linguistics.

References

1. Bender, E.M.; Gebru, T.; McMillan-Major, A.; Shmitchell, S. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 2021, pp. 610–623.
2. Seaman, B. Generative Works: From Recombinant Poetics to Recombinant Informatics. 2014 International Conference on Cyberworlds, 2014, pp. 5–11. doi:10.1109/CW.2014.10.
3. Lenat, D. Sometimes the Veneer of Intelligence is Not Enough. *Cognitive World* **2018**.
4. Seaman, B. Towards A Dynamic Heterarchical Ecology Of Conversations. *Heinz von Foerster Lecture* **2017**.
5. Seaman, B. Neosentience and the Abstraction of Abstraction. *Systems. Connecting Matter, Life, Culture and Technology* **2013**, *1*, 51.
6. Givón, T. *Mind, code and context: Essays in pragmatics*; Psychology Press, 2014.
7. Minsky, M. A Framework for Representing Knowledge. In *Mind Design 2: Philosophy, Psychology, Artificial Intelligence*; Haugeland, J., Ed.; MIT Press: Cambridge, 1997; pp. 111–142.
8. Weyhrauch, R.W. Prolegomena to a Theory of Mechanized Formal Reasoning. *Artificial Intelligence* **1980**, *13*, 133–170.
9. Weyhrauch, R. Ideas on Building Conscious Artifacts. *The FOL Project* **1991**.
10. Weston, J.; Bordes, A.; Chopra, S.; Rush, A.M.; van Merriënboer, B.; Joulin, A.; Mikolov, T. Towards Ai-Complete Question Answering: A Set of Prerequisite Toy Tasks. *ArXiv Preprint ArXiv:1502.05698* **2015**.
11. Searle, J.R. Indirect speech acts. In *Speech acts*; Brill, 1975; pp. 59–82.
12. El Asri, L. Talking with Machines with Dr. Layla El Asri. *Microsoft Research Podcast* **2019**.
13. Perrault, C.R.; Allen, J. Speech Acts as a Basis for Understanding Dialogue Coherence. Theoretical Issues in Natural Language Processing-2, 1978.
14. Appelt, D.; Konolige, K. A Practical Nonmonotonic Theory for Reasoning About Speech Acts. 26th Annual Meeting of the Association for Computational Linguistics, 1988, pp. 170–178.
15. Cohen, P.R.; Perrault, C.R. Elements of a Plan-Based Theory of Speech Acts. *Cognitive Science* **1979**, *3*, 177–212.
16. Allen, J.; Hinkelman, E.A. Using Structural Constraints for Speech Act Interpretation. Speech and Natural Language: Proceedings of a Workshop Held at Cape Cod, Massachusetts, October 15-18, 1989, 1989.
17. Nelson, K. *Narratives from the Crib*; Harvard University Press: Cambridge, 2006.
18. Turner, M. *The origin of ideas: Blending, creativity, and the human spark*; Oxford University Press, 2014.
19. Adam, A. *Artificial Knowing: Gender and the Thinking Machine*; Routledge, 2006.
20. Lamoureaux, S.; Hagerty, A. Women, Machines, and Dangerous Things: Animating Intelligent Personal Assistants as Semantico-Pragmatic Violence. In *The Gender of Things: How Technologies and Epistemic Objects Become Gendered*; Rentetzi, M.; Bosch, A., Eds.; Routledge, Forthcoming.