Applying Deep Language Understanding to Open Text: Lessons Learned

Marjorie McShane (margemc34@gmail.com)
Stephen Beale (stephenbeale42@gmail.com)
Irene Nirenburg (irene_bn@yahoo.com)

Cognitive Science Department, Rensselaer Polytechnic Institute
Troy, NY 12180 USA

Abstract

Human-level natural language understanding (NLU) of open text is far beyond the current state of the art. In practice, if deep NLU is attempted at all, it is within narrow domains. We report a program of R&D on cognitively modeled NLU that works toward depth and breadth of processing simultaneously. The current contribution describes lessons learned – scientifically and methodologically – from an exercise in applying deep NLU to open-domain texts. An overarching lesson was that although learning to compute sentence-level semantics seems like a natural step toward computing full, context-sensitive, semantic and pragmatic meaning, corpus evidence underscores just how infrequently semantics can be cleanly separated from pragmatics. We conclude that a more comprehensive methodology for automatic example selection and result validation is needed as prerequisite for success in developing NLU applications operating on open text.

Keywords: natural language understanding; cognitive modeling; language-endowed intelligent agents

Introduction

Operationalizing human-level natural language understanding (NLU) in computer systems has been a goal of AI since its inception. People want intelligent agents to understand not only what they say but what they mean, taking into account the linguistic and real-world context, shared background knowledge, the interlocutors' mutually understood plans and goals, and even their mental, physical, and emotional states. All of these considerations explain why human-level NLU is an AI-complete problem.

It is difficult to carve out a program of R&D for AI-complete problems. With respect to natural language, the field has responded in five broadly-defined ways.[1] **(1)** *Avoid meaning.* For the past 25 years, mainstream NLP has chosen to pursue so-called *knowledge-lean* methods, i.e., the statistical processing of big data with little to no computation of meaning. This has proven useful for certain applications but is not moving toward explainable, human-level NLU in service of intelligent agents. **(2)** *Address select aspects of meaning.* Computing individual aspects of meaning has improved the quality of some primarily knowledge-lean systems. Topics addressed include, e.g., case-role identification, speech act detection, textual coreference resolution, and the semantic clustering of word strings using distributional semantics (Jurafsky and Martin, 2009). **(3)** *Pursue deep NLU in a (very) narrow domain.* This provides systems with the kinds of knowledge and reasoning capabilities that people leverage when interpreting language (e.g., Allen et al. 2007; Lindes and Laird, 2016). **(4)** *Build theories without systems.* Such work anticipates that prerequisites – such as NLU – will be eventually be fulfilled externally, and is typical in the fields like computational formal semantics and machine reasoning. **(5)** *Build extensive theories but implement and evaluate just a subset*. This appears to be the choice of the dialog specialist David Traum (compare Traum 1994 for scientific work with Nouri et al. 2011 for application-oriented work).

The program of R&D described here – developing Language-Endowed Intelligent Agents (LEIAs) within the OntoAgent cognitive architecture – offers a sixth approach to attacking the AI-complete problem of human-level NLU (McShane, Nirenburg and English, 2018). It pursues **depth of analysis and breadth of coverage concurrently**, but with appropriately flexible expectations about the coverage, quality, and confidence of analyses depending on the correlation of text inputs with knowledge bases. It focuses on the *actionability* of language interpretations, as judged by the agent systems that use them.

Of the many theoretical and methodological issues at the core of this program of work (McShane, Nirenburg and Beale, 2016), the following are particularly relevant for this discussion.

1. LEIAs are modeled after humans. Like humans, they do not need to understand everything their interlocutors say and mean; instead, they need to achieve actionable interpretations, defined as interpretations that are sufficient to support reasoning about action.
2. The same knowledge that allows LEIAs to function intelligently in their domain of expertise supports language-oriented reasoning in that domain. Full NLU is not possible without such knowledge.

[1] This is a thumbnail sketch of a long history and extensive literature. See Nirenburg and McShane (2016) for a more in-depth treatment.

3. For both theoretical and methodological reasons, NLU is best implemented as a series of layers of ever-deeper analysis, resulting in ontologically-grounded text meaning representations (TMRs) that are well-suited to agent reasoning.
4. Most narrow-domain approaches seek to avoid disambiguation, one of the hardest problems of NLU; however, such approaches will not attain a human level of understanding until this problem is solved and agents function with a realistic-sized lexicon.
5. A very large number of linguistic phenomena (to name just a few: nominal compounding; all aspects of reference resolution, including fragments and ellipsis; non-literal language; indirect modification; indirect speech acts; implicatures) must be handled by LEIAs no matter their domain of specialization (McShane and Nirenburg, *forthcoming*). The computational microtheories accounting for these phenomena are best investigated using open text.

The original hypothesis underlying the work described here was that we could *quickly* validate many of the implemented microtheories of NLU for LEIAs using an open corpus. Why an open corpus? As discussed in more detail later, this method a) provides useful fodder for improving microtheories, b) makes the work "real" in the eyes of the mainstream NLP community, and c) shows how the analysis capabilities can be usefully applied to open texts.

In formulating the reported exercise, we assumed that a corpus would contain a sufficient number of sentences that could be automatically interpreted using general linguistic and world knowledge, without the need for the finer-grain knowledge resources supporting agent-reasoning capabilities that are available only in narrower domains. Such sentences would be similar in nature, but methodologically preferable, to the invented examples we use to test out individual microtheories.

We further assumed that a simple, automatic method of extracting examples would serve the purpose. However, this experience has shown that, in order to sufficiently evaluate all of the microtheories contributing to the system, we need a more sophisticated example extraction methodology operating over a larger corpus, as well as more human effort devoted to reviewing results. However, rather than change the original hypothesis by allocating more time and effort to data collection, we heeded the lessons learned from the Reproducibility Project (Open Science Collaboration, 2015) and its analytical wake: It is not appropriate to tweak hypotheses or results until they achieve the envisioned threshold. Research habitually involves things not going to plan, and the associated lessons learned are central to progress in the field. This paper focuses on lessons learned. But we must begin with the briefest introduction to the NLU environment at hand.

The OntoAgent Cognitive Architecture

The OntoAgent cognitive architecture underlying LEIAs includes the modules of perception, reasoning and action. Language is one of the perception modes of a LEIA. Language inputs are analyzed into disambiguated, ontologically-grounded meaning representations. For example, the bare-bones basic TMR for *I knocked on the door* (stripped of metadata and calls to the procedural semantic routines for coreference resolution) is as follows:

```
(HIT-1 (AGENT HUMAN-1)
       (THEME DOOR-1)
       (INSTRUMENT HAND (OPENNESS 0))
       (TIME < find-anchor-time)) ; indicates past tense
```

The fact that the instrument is a closed hand is provided by the lexical description of the selected sense of *knock* in the system's lexicon, which also expects the object of the preposition to refer to, among other possibilities, a door.

Although we cannot adequately familiarize readers with the theory of Ontological Semantics, the agent applications that this approach to NLU has supported, the knowledge bases employed, or how the analysis process works (see, e.g., Nirenburg and Raskin, 2004; McShane, Nirenburg, and English, 2018; Nirenburg, McShane and Beale, 2008), the following facts will serve as orientation. The lexicon contains ~30,000 word senses, which are comprised of linked syntactic and semantic representations and, whenever necessary, calls to procedural semantic routines (for example, to resolve coreferences). Argument-taking words, multiword expressions, and polysemy are richly represented. The semantic descriptions are written in an unambiguous ontological metalanguage. The ontology contains ~9,000 concepts (~145,000 RDF triples), mostly from the general domain. Concepts are described using attributes and relations. Scripts detailing complex events are available in select domains.

The lexicon and ontology were mostly compiled through a modest, short-term effort around 25 years ago in service of interligua-based machine translation and have been only minimally modified since. They were not modified at all for the reported exercise. The key benefit of our lexicon is that it is far from toy and, therefore, allows us to develop and test the essential capability of lexical disambiguation. All parts of speech include polysemous entries, and light verbs such as *have, make,* and *do* have dozens of senses, many of which involve multi-word expressions or constructions. The ontology, for its part, provides selectional constraints on case-roles that support disambiguation, as well as a substrate for various types of language-oriented reasoning, such as topic/domain detection based on ontological distance.

Although these resources have served our research goals quite well, their insufficiencies are relevant to the current report. We estimate that the lexicon would need to be around ten times larger to provide baseline coverage of open text, with the necessary acquisition including a large percentage of multi-word expressions and constructions. An

acquisition effort of this size is, we estimate, no more labor-intensive than some of the well-known corpus annotation efforts in service of supervised machine learning.

NLU by LEIAs is reasoning-intensive. The overall process is modeled as two types of incrementality: *horizontal incrementality* involves analyzing elements of input as they become available to the agent (essentially, word by word); *vertical incrementality* involves applying, on an as-needed basis, increasingly sophisticated methods of analysis to the given state of input, be it a fragment, a complete utterance, or a multi-sentence text. Agents dynamically decide how deeply to process chunks of input as they are perceived.

There are six stages of vertical incrementality, described in greater detail in (McShane and Nirenburg, *forthcoming*): **1.** Preprocessing and syntactic parsing, for which we use the CoreNLP toolset (Manning et al. 2014). **2.** Integrating these results into our environment, which includes recovering from unexpected syntax as well as the initial stage of learning new words. **3.** Basic semantic analysis, which uses lexical and ontological knowledge for disambiguation and semantic dependency analysis. This includes such advanced capabilities as the detection and resolution of many types of ellipsis and learning the semantics of unknown words. **4.** Aspects of reference resolution that do not require full contextual grounding. These include resolving textual coreference, identifying which referring expressions do not require a coreferent and why, establishing reference relations that are not *co*reference (e.g., bridging constructions), and reconsidering upstream lexical disambiguation decisions based on coreference relations. **5.** Extended semantic analysis, which treats select instances of residual ambiguities and incongruities using additional general-purpose rule sets. These include, e.g., ontological patterns for interpreting nominal compounds, rules for interpreting metonymies, and dialog-analysis strategies for integrating the meaning of fragmentary utterances into the discourse. **6.** Situated NLU, which applies all of an agent's domain-specific and situational knowledge and reasoning to resolve residual ambiguities and incongruities, and anchors newly learned knowledge to agent memory.

If it sounds like this system is claiming to do *everything,* that is, in a certain sense, correct. The overall challenges of NLU must be addressed in an integrated system, within an architecture and theory that reserves a place for each component microtheory. The microtheories must be crafted as components of such an overall analysis system. This approach avoids the two most serious problems of strictly modular or limited-scope research: the assumption that prerequisites for one's own work will be provided externally; and the avoidance of all cross-modular phenomena.

Stages 1-5 involve what some call *semantic* meaning, as contrasted with *pragmatic* (discourse, situational) meaning. This level of meaning should be understandable at the sentence level, outside of context – even if some expressions (e.g., pronouns) remain underspecified. Following this expectation, individual sentences outside of their context were the focus of the reported exercise. Given that the ~30,000-sense lexicon contains over 1,600 verb senses, and that the system can process proper nouns and learn new words, we projected that there would be plenty of appropriate sentences to seed our exercise. As concerns Stage 6 of processing, it cannot be validated using individual sentences outside context; we are working on that separately, within a robotic application (Nirenburg et al., 2018).

Methodology

Our initial goal was to focus on *validating* our system rather than formally *evaluating* it in the way that has become standard in the field of natural language processing (NLP). That methodology is of no use for systems that seek human-level understanding of language. It is not, therefore, surprising that mainstream NLP has all but officially placed our area of R&D beyond the boundaries of the discipline. For example, in their chapter on "Evaluation of NLP Systems" in *The Handbook of Computational Linguistics and Natural Language Processing* (Clark, Fox and Lappin, 2010), Resnik and Lin do not even address the evaluation of cognitively-oriented systems that integrate scientific and technological goals. They write: "such scientific criteria [involving, e.g., the cognitive modeling of human language processing] have fallen out of mainstream computational linguistics almost entirely in recent years in favor of a focus on practical applications, and we will not consider them further here." (p. 271) So, we need an alternative validation/evaluation methodology.

There is no truly fast, easy, and complete way to validate (no less evaluate) a large and complex knowledge-based system, nor can the full set of options be fleshed out in this short space. As a starting point, consider just a few of the options. **(1)** *Invent test inputs guided by the knowledge bases and system capabilities*. This gives credit for what *does* work but rarely uncovers unexpected phenomena and is viewed skeptically by the field at large. **(2)** *Use inputs limited to a narrowly-defined domain.* This, too, usually involves manual example creation since 'narrowly-defined' must be enforced; moreover, it fails to give the system or component microtheories credit for their applicability across domains. **(3)** *Use randomly selected inputs from the open domain*. Although this is a cornerstone of statistical NLP, it is inapplicable to deep NLU given that the environment is known to have limited lexical coverage. **(4)** *Focus on full sentences from open text that the system analyzes perfectly.* This approach tasks the system with extracting from open text, and processing, only those sentences it hypothesizes it can analyze correctly. During validation, people inspect only the highest-quality results – i.e., those for which exactly one TMR achieves the highest score, and that score reflects high confidence. This is the approach we used for the current exercise. Its insufficiencies underlie many of the lessons learned from this exercise, as discussed in the next section. **(5)** *Focus on subsentential chunks of text from the open domain that the system analyzes perfectly.* Such chunks can represent propositions, individual phenomena (e.g., nominal compounds, instances of verb phrase ellipsis),

or sentences for which all aspects but one – e.g., an unknown adverb – are correctly understood. We have used this method in past formal evaluations (Nirenburg et al., 2018) and have found it useful for vetting individual microtheories. The problem is that it is time-consuming to formulate a vetting regimen for even a single microtheory, let alone the dozens that the system currently comprises, or the interactions among them. Additionally, any of the above methods can also involve inspecting outputs that are partially correct, residually ambiguous, etc.

Results

As should be clear by now, this vetting exercise was primarily intended to guide our continued R&D effort. It did – but more through lessons learned than from compiling examples that work. That being said, we do want to present some examples to show that our NLU system *can,* in fact, work on open text.

To further specify the set-up: The system extracted examples from two randomly selected excerpts of the COCA corpus (Davies 2008), one literary and the other journalistic. It extracted sentences that included a maximum of one unknown word, with "known" implying that the lexicon contained an entry with the necessary part of speech. No other extraction filters were applied. The system processed the sentences into TMRs using Stages 1-5 of our NLU system. We manually reviewed only those results that seemed promising. For example, we did not inspect the TMRs for sentences that were incomprehensible outside of context, or that required knowledge or reasoning beyond that available in Stages 1-5 of NLU.

We spent just a few person-weeks on the exercise, much of which involved code debugging (after all, the exercise was primarily in service of R&D). However, the examples we cite as "correct" were correct *before* any system modifications. No amendments to the knowledge bases were made. It did not take long to determine that we had learned what we could from this exercise, and we, therefore, did not prolong it to collect more working examples.

Unless otherwise noted, all examples presented in this section were analyzed perfectly. Any incorrect portions are indicated by strikethroughs or explanatory text. Every input required disambiguation decisions, in some cases, from a large choice space: e.g., *He looked for the creek* disambiguates between 16 senses of *look*, and *I went into the bathroom* disambiguates between 54 senses of *go*. The examples below are grouped by the specific phenomena they illustrate.

Complex semantic descriptions. For example, the TMR for *I knocked on the door* includes a hand as the instrument, and the TMR for *I pointed at the blood* includes a finger as the instrument.

Disambiguation of highly polysemous particles and prepositions: *She rebelled against him; He stared at the ceiling; She jokes with him; She switched on the light; He passed through the entrance; I called for a blanket; I thought about Amalia; He talked about Leona***.**

Modification and sets: *An old white couple lived in a trailer.*

Multiword expressions: *He took me by surprise*,

Verbal disambiguation using a specificity preference. For example, in *I do not know Dave,* three senses of *know* (glossed as *be acquainted with, be aware of,* and *be able to identify*) formally match the case-role constraints. The sense *be acquainted with* fulfills the tightest case-role constraints, so it wins. This example also shows the correct processing of the modality indicated by negation.

Dynamic sense bunching. This allows the system to underspecify an interpretation rather than end up with competing analyses. E.g., ~~*No, and*~~ *I didn't ask him* does not permit disambiguation between three senses of *ask* – those encoded using the ontological concepts REQUEST-INFO, REQUEST-ACTION and PROPOSE – so the system bunches these into their closest common ontological ancestor, ROGATIVE-ACT, whose case-roles are correctly understood as AGENT and THEME.

Lateral selectional constraints for disambiguation. E.g., in *I heard the hands on the clock* ~~*move*~~*, clock* was correctly used to disambiguate *hands* (but since the CoreNLP misidentified "clock move" as a nominal compound, that aspect of the analysis was wrong). Similarly, in *The arm jerked, eyelids* ~~*rose*~~*,* the meaning of *eyelids* was correctly used to disambiguate *arm* between body part and furniture part (but *rise* as applied to eyelids was misanalyzed).

New word learning. An example of new noun learning is 'uncle' in *The uncle said something to him*, which is understood as referring to a HUMAN since the AGENT slot of ASSERTIVE-ACT must be filled by a HUMAN. The results of learning are understood as provisional, and values of properties of the newly learned concept are expected to be added opportunistically as a side effect of continued processing of input – or, alternatively, by a knowledge acquirer. An example of new property learning is *inconsiderate* in *Burying Leora* ~~*in Pittsburgh*~~ *is inconsiderate*. The system represents the meaning as a generic PROPERTY whose DOMAIN is filled by the event BURY (from *burying*). *In Pittsburgh* was correctly analyzed but incorrectly attached to Leora rather than burying, following a parsing error by CoreNLP. (Reambiguating PP attachments from the CoreNLP parse, so that semantics can weigh in, is on agenda.)

The above presents just a small sampling of linguistic phenomena that the system covers, along with examples of successful analyses. It shows that vision behind the current exercise was not ultimately ill-conceived, and illustrates that the corpus was, in fact, open-domain. But, as we said earlier, we keep this aspect of the report brief in order to focus on the main point: lessons learned.

Lessons Learned

Most of the *types* of outcomes of this exercise were predicable beforehand, but in some cases their *frequency* was rather surprising, thus representing a lesson learned.

1. *It is not possible to automatically detect that a needed multiword expression (idiom, construction, etc.) is missing*

in the lexicon. Multiword expressions are central to a human's knowledge of language and, accordingly, to modeling NLU for LEIAs. When a multiword expression is missing from the lexicon, the system analyzes the components compositionally, which necessarily results in an error. All of the following examples were misanalysed because interpreting the meaning of the underlined portion required a multiword lexical sense that had not yet been acquired. *She is long gone from the club. I got a good look at that shot; The Knicks can live with that. But once Miller gets on a roll, he can make shots from almost 30 feet. I can't say enough about him. This better be good. You miss the point. I should have known better.* The lesson learned involves the frequency with which the system will be overly confident in its analysis, not having recognized that an input component is not semantically compositional.

2. *The methodology of focusing on completely correct TMRs was suboptimal.* Often, the meaning representation of a portion of the input nicely demonstrates a particular functionality, even though some aspect of the overall sentence interpretation is incorrect. Many such mistakes reflect the use of microtheories that are currently underdeveloped, such as those for relative temporal and spatial relations (*in recent weeks, 25 feet right of the hole*, and *for the second time this year*). When, midstream, we decided to revisit partially correct TMRs, we found many interesting correct subanalyses, suggesting that vetting Method #5 described above might be superior to the method we used.

3. *The methodology of focusing exclusively on sentences that resulted in a single TMR was suboptimal*. Outside of context, residual ambiguity is quite common. When we decided to revisit analyses that resulted in two output TMRs – because the analyzer did not have a reason to prefer one over the other – we found examples in which this outcome was actually the correct one. For example, the system correctly detected the ambiguity, and generated multiple correct candidates, for *He stared at the fish,* which could refer to a live fish (FISH) or its meat (FISH-MEAT); and *He glanced at the walls* could refer to parts of a room (WALL) or parts of a person undergoing surgery (WALL-OF-ORGAN).

4. *It can be difficult, even for humans, to describe many intended meanings.* Consider the following sentences: *And he came back from the dead. Training was a way of killing myself without dying. The supporting actor has become the leading man. This is about substance. The roots that are set here grow deep.* Such examples allow for multiple interpretations, at many levels of vagueness and specificity, depending on the specific speech situation. The existence of utterances of this type are among the reasons we believe that, in building agent-oriented NLU capabilities, actionability – not exhaustive understanding – is key. But for this exercise, decision-making about actionability was outside of purview.

5. *The intended meaning can rely more centrally on discourse/pragmatic interpretation than semantic analysis*. In some cases, e.g., for personal pronouns, there is a clear progression from semantic to pragmatic meaning. However, in other cases, semantic meaning is either vague, not directly connected with pragmatic meaning, or even relatively unimportant. Space is too short to flesh out these complex eventualities, but consider the example *It takes two to tango*, which occurred in our corpus. If we were to write a lexical sense for this phrase, how would we describe its meaning? Its propositional meaning – something like "a communication cannot exist without multiple people being agentive" – is much less important than its discourse function. That is, the speaker is saying that the given situation is an example of a generalization about human relations, but the context-specific pragmatic nuances can range from being a barb during a spat (*It's your fault, too, that we're arguing!*) to being advice to a friend (*If you back off, maybe the other person will too*). It seems incorrect to lexically record, and then give a system credit for computing, semantic meanings when it is the pragmatic force that is predictably more important.

6. *Non-literal language is even more prevalent than we had expected – and we had expected a lot.* In fact, we have methods for detecting and recovering from some types of non-literal language, but not the onslaught we encountered in this exercise. For example, *Everyone was saying we won ugly last week* and *He not only hit the ball, he hammered* were imperfectly analyzed because the non-literal meanings were not correctly recovered.

7. *We need to operationalize reasoning about language via affordances*. Just as human vision is well-understood to be largely driven by expectations, so, too, is language understanding. Affordances – i.e., the knowledge of what objects can do and how they can be used – can support reasoning about language inputs, particularly if they involve difficult phenomena, such as non-literal language, unknown words, and indirect modifications. For example, we previously noted that *eyelids rose* resulted in a misinterpretation of 'rise'. It is unlikely that people encode a word sense of 'rise' that covers eyelids; however, we know that eyelids are capable of precious few actions. So a fuzzy matching between words and concepts for moving up and down is sufficient for a person to understand this. A microtheory of applying affordances to reasoning about NLU is on our team's agenda.

8. *It is unclear what credit to give semantics without implicatures*. On the one hand, semantic analysis is hard enough without requiring that NLU systems account for all a speaker's implicatures before claiming any success. On the other hand, in some cases semantics and implicatures cannot be neatly separated. Consider the example, *She's also a woman*. Reading this in isolation, we understand that the context must have been about her in some other social role – as a mother, a co-worker, etc. – and that this utterance focuses attention on her female/sexual side. It is similarly unclear what, if anything, would count as a sufficient semantic (pre-implicature) analysis of the following: *How quickly the city claimed the young. They sat by bloodline. I think he is coming into good years. Fathers were for that.*

9. *Not invoking domain-oriented expectations is more limiting that we had anticipated.* For example, unless you

realize you are in a sports context – and know sports-related lexical and ontological knowledge – the following are not fully interpretable: *The Rangers and the Athletics have yet to make it. He hit his shot to four feet at the 16th. We stole this one. I wanted the shot.*

10. *Although our system is knowledge-based and all processing apart from what is contributed by CoreNLP is fully inspectable, the computational complexity of deep NLU can make it difficult to fully predict, explain, and trouble-shoot results*. Consider the simple example *I almost never talk about it,* whose words have, respectively, 3/5/2/2/2/1 senses. The number of candidate TMRs generated is 50, with their final scores ranging from -22 to 22.9. There was only 1 highest-scoring TMR and it was correct. The CoreNLP parse happened to be correct, but since this is not always the case, our analyzer compensates by considering other syntactic analysis possibilities as well. As a result, the process of mapping syntactic dependencies in inputs to the variables in the syntactic descriptions in lexicon entries can lead to multiple sets of variable assignments for each available sense. At the semantic level, the system needs to select the best sense and set of variable assignments for each word by examining the interactions between the semantic constraints among all the words that interact with it. In the worst case, that can become a computational clique, which has exponential time requirements (we employ various techniques for reducing or sometimes eliminating this computational drain). Scoring functions are also complex – their composition is a research issue in itself. In short, even though we can configure a glassbox evaluation, the analysis process can, in certain cases, still defy complete explanation.

Conclusions

We believe that our original goal – to vet our system's domain-independent microtheories using open text – is achievable. The reason why the focus of this exercise shifted from "vetting" to "investigating lessons learned" is because the methodology for extracting examples and automatically evaluating the quality of output TMRs turned out to be insufficiently developed. The lessons learned will inform the creation of a more sophisticated methodology for future experiments. To give just a few examples of planned enhancements: (a) Including the preceding context for each extracted example to allow for coreference and lateral-constraint heuristics to be leveraged; (b) Automatically excluding excessively short inputs, direct speech, texts from jargon-intensive domains like sports, and inputs containing pronouns whose resolution strongly affects disambiguation decisions (e.g., *it, that* and *they* are more problematic than *he* or *she*); (c) Using an example-extraction methodology that identifies the highest-confidence examples of each word sense, microtheory, etc., from a much larger corpus than was used for this exercise; and (d) Including within purview high-confidence subsentential results.

Apart from lessons learned, this experiment has resulted in promising outcomes. The fact that the system correctly analyzed some inputs from the open domain – even given the shortcomings of the reported methodology and all of the challenges natural language predictably presents – suggests that deep NLU can have near- and mid-term utility, given an appropriate task formulation and improved methods of automatically judging the system's confidence in its analyses.

Lifelong learning has long been understood as a necessary foundation of AI. Even the current capabilities of the reported NLU system can support the learning of lexical units and ontological concepts, with the coverage expected to rise dramatically even with relatively (by industry standards) modest knowledge acquisition efforts.

Our system addresses the open-world problem directly and takes responsibility for all upstream processing errors (currently, from CoreNLP). In some cases, it can successfully learn new meanings and recover from upstream errors, whereas in others it cannot. However, we believe that failures under real-world circumstances are far preferable to the non-real-world experimental set-ups favored by the well-known task-oriented competitions of statistical NLP.

Although the reported exercise focused on stages of NLU that can be, to some degree, computed outside of context, the overall program of work moves toward explainable AI covering integrated agent functionalities.

Acknowledgments

This research was supported in part by Grants #N00014-16-1-2118 and # N00014-17-1-2218 from the U.S. Office of Naval Research. Any opinions or findings expressed in this material are those of the author and do not necessarily reflect the views of the Office of Naval Research.

References

Allen, J. F., Chambers, N. et al. (2007). PLOW: A collaborative task learning agent. *Proceedings of the 22nd National Conference on Artificial intelligence* (AAAI '07), Vol. 2, pp. 1514-1519. AAAI Press.

Davies, M. (2008-). The Corpus of Contemporary American English: 450 million words, 1990-present.

Clark, A., Fox, C., & Lappin, S. (2010). *The Handbook of Computational Linguistics and Natural Language Processing*. Wiley-Blackwell.

Jurafsky, D., & Martin, J. H. (2009). *Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics*. 2nd edition. Prentice-Hall.

Lindes, P., & Laird, J. E. (2016). Toward integrating cognitive linguistics and cognitive language processing. *Proceedings of the 14th International Conference on Cognitive Modeling* (ICCM). University Park, Pennsylvania.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., & McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations,* pp. 55–60. Stroudsburg, PA: The Association for Computational Linguistics.

McShane, M., & Nirenburg, S. (Forthcoming). Context for language understanding by intelligent agents. *Applied Ontology*.

McShane, M., Nirenburg, S., & Beale, S. (2016). Language understanding with Ontological Semantics. *Advances in Cognitive Systems* 4: 35-55.

McShane, M., Nirenburg, S., & English, J. (2018). Multi-stage language understanding and actionability. *Advances in Cognitive Systems* 6: 1-20.

Nirenburg, S., & McShane, M. (2016). Natural language processing. In Chipman, S. (Ed.), *The Oxford Handbook of Cognitive Science, Volume 1*. New York: Oxford University Press. Online publication date: August 2016.

Nirenburg, S., McShane, M., & Beale, S. (2008). A simulated physiological/cognitive "double agent." In Beal, J., Bello, P., Cassimatis, N., Coen, M. & Winston, P. (Eds.), *Papers from the Association for the Advancement of Artificial Intelligence (AAAI) Fall Symposium "Naturally Inspired Cognitive Architectures*". AAAI Technical Report FS-08-06. Menlo Park, CA: AAAI Press.

Nirenburg, S., McShane, M., Beale, S., Wood, P., Scassellati, B., Mangin, O., & Roncone, A. (2018). Toward human-like robot learning. *Natural Language Processing and Information Systems,* Proceedings of the 23rd International Conference on Applications of Natural Language to Information Systems (NLDB 2018), pp. 23-82.

Nirenburg, S., & Raskin, V. (2004). *Ontological Semantics*. The MIT Press.

Nouri, E., Artstein, R., Leuski, A., & Traum, D. (2011). Augmenting conversational characters with generated question-answer pairs. *Proceedings of the AAAI symposium on question generation*.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), aac4716. Doi: 10.1126/science.aac4716

Traum, D. R. (1994). A Computational Theory of Grounding in Natural Language Conversation. PhD thesis, Department of Computer Science, University of Rochester.