Self-Confidence in Autonomous Systems
Papers from the AAAI 2015 Fall Symposium

OntoAgents Gauge Their Confidence In Language Understanding

Marjorie McShane and Sergei Nirenburg
Rensselaer Polytechnic Institute, Troy, NY
margemc34@gmail.com, zavedomo@gmail.com

Abstract

This paper details how OntoAgents, language-endowed intelligent agents developed in the OntoAgent framework, assess their confidence in understanding language inputs. It presents scoring heuristics for the following subtasks of natural language understanding: lexical disambiguation and the establishment of semantic dependencies; reference resolution; nominal compounding; the treatment of fragments; and the interpretation of indirect speech acts. The scoring of confidence in individual linguistic subtasks is a prerequisite for computing the overall confidence in the understanding of an utterance. This, in turn, is a prerequisite for the agent's deciding how to act upon that level of understanding.

Introduction

The concept of self-confidence, as applied to intelligent agents, is rather broad: it clearly can be applied to any and all decisions that an agent must make in carrying out its tasks, whether they relate to perception, reasoning and decision-making, or action. Before self-confidence can be used to inform decision-making, it must first be assessed.

In this paper we focus on the assessment of self-confidence with respect to language understanding. Specifically, we discuss metrics for establishing self-confidence with respect to the following processes that contribute to the agent's overall language understanding task: lexical disambiguation and the establishment of semantic dependencies; reference resolution; nominal compounding; the treatment of fragments; and the interpretation of indirect speech acts. For each of these tasks, we propose an inventory of salient features that contribute to the calculation of confidence. The confidence levels for individual decisions provide evidence for the construction and optimization of the agent's overall confidence in its understanding of an input.

A fine grain size of heuristic feature specification like the one presented here is only required if the overall objective of research is building explanatory models of agency. In the long term, this approach is a more promising path toward having artificial agents approach human levels of performance in everyday functioning than the alternative of using purely predictive (non-explanatory) models.[1] Our work combines building explanatory models of agency with working on practical applications for which our models might be able to have an impact even in the near-to-mid-term (e.g., the Maryland Virtual Patient prototype application; Nirenburg et al., 2008).

In this paper, we present an inventory of heuristic features for each of the language understanding tasks listed above, as well as our methodology for formalizing the scoring metrics, which is work in progress. The paper is organized as follows. First we very briefly describe the OntoSem approach to language analysis, which is used by OntoAgents. Then we present language analysis tasks that an agent must perform in order to arrive at an interpretation of a language input. For each task we describe an inventory of metrics for assessing confidence in the results of the associated processing. Since our approach to language understanding is largely human-inspired – i.e., we undertake to model the process of human language understanding, not only emulate its results – so, too, are most of the metrics for judging confidence. We conclude with a brief assessment of the work and directions of ongoing research.

Language Understanding In OntoAgent

The goal of OntoSem language understanding is to automatically generate unambiguous, fully specified, ontologically-grounded text meaning representations (TMRs), as defined by the theory of Ontological Semantics (Nirenburg

[1] By contrast, predictive models currently offer the best price/quality ratio for building application systems of immediate practical utility.

and Raskin, 2004; McShane et al., forthcoming-a). TMRs are stored in agent memory and serve as comprehensive and unambiguous inputs to reasoning (McShane and Nirenburg, 2012). For example, the TMR for the input *You need to apply pressure to the wound* is as follows (excluding metadata):

```
REQUEST-ACTION-1
   AGENT          HUMAN-1    ; the speaker
   THEME          PRESS-1
   BENEFICIARY    HUMAN-2    ; the interlocutor

PRESS-1
   AGENT          HUMAN-2
   THEME          WOUND-INJURY-1
```

The TMR is headed by a numbered instance of the concept REQUEST-ACTION. The AGENT of this action is the HUMAN speaker and its THEME (what is requested) is a PRESS event. The PRESS event is further specified, in its own frame, as having the HUMAN interlocutor as its AGENT and a WOUND-INJURY as its THEME.

Language analysis is supported by a 30,000-sense lexicon that includes linked syntactic and semantic descriptions of the meanings of words and phrases. For example, the multi-word expression *apply pressure to* is recorded as a sense of the verb *apply* (for more on multi-word expressions, see McShane et al., forthcoming-b):

```
apply-v5
    def     "phrasal: apply pressure to"
    ex      "She applied pressure to his chest."
  syn-struc
    subject         (root $var1) (cat np)
    v               (root $var0) (cat v)
    directobject    (root $var2) (cat n) (root pressure)
    pp              (root $var3) (cat prep) (root to)
                    (obj ((root $var4) (cat np)))
  sem-struc
    PRESS
      AGENT     ^$var1
      THEME     ^$var4
    ^$var2    null-sem+
    ^$var3    null-sem+
```

The syn-struc asserts that, in the active voice, we expect a subject, the verb *apply*, the direct object *pressure* and a prepositional phrase headed by the preposition *to*. Each of these elements is associated with a variable. The sem-struc asserts that the event in question is an instance of the ontological concept PRESS, whose agent is the meaning of $var1 (^ indicates "the meaning of") and whose THEME is the meaning of $var4. The ontological description of PRESS – which is consulted during lexical disambiguation – includes the information that the AGENT must be ANIMATE and the THEME must be a PHYSICAL-OBJECT that is NOT ANIMATE. (So, an input like *He applied pressure to his employees* will be treated by a different lexical sense whose prepositional object is constrained to HUMAN.) The descriptor "null-sem+" indicates that the meaning of these elements should not be computed compositionally – it has already been taken care of. The TMRs that result from text analysis are stored to agent memory, to serve as input to reasoning.

Lexical Disambiguation and Semantic Dependency Determination

The above TMR shows the results of lexical disambiguation and the establishment of the semantic dependency structure. To create TMRs, the system considers every lexical sense of every word and attempts to optimize the alignment between the expectations of argument-taking senses in the lexicon and the available interpretations of elements in the input. If an input precisely aligns – syntactically and semantically – with a given selection of lexical senses, then the interpretation will receive a high score. For *apply-v5* this occurs, for example, with the input *She applied pressure to his chest*. The expected syntactic constituents occur precisely as expected: *subject "apply" direct-object "to" object-of-preposition.* The semantic constraints also match perfectly: in the ontological specification for PRESS, the AGENT is constrained to ANIMAL (a superclass of HUMAN) and the THEME is constrained to PHYSICAL-OBJECT (a superclass of CHEST-BODY-PART). If only one of the verb's senses aligns with the input in this way, and if the system is confident in its disambiguation decisions for the verb's arguments (a separate calculation altogether, which often involves reference resolution), then the task of lexical disambiguation and semantic dependency determination receives a very high score.

However, other contexts can include features that affect confidence levels either positively or negatively. In what follows, we briefly describe a subset of these features.. Note that we do not at this time address the relative importance of each of the features for the overall cumulative confidence level. That is a task for future work. At present we concentrate on the qualitative theoretical task of formulating the individual heuristic features associated with core language phenomena and the methods of determining their effect on the agent's confidence in its choice of solutions.

1. The input matches > 1 lexical sense, featuring looser and tighter semantic constraints (incurs a penalty).

Sometimes more than one lexical sense perfectly matches an input. Typically, some of the senses have tighter selectional constraints whereas others have broader ones. For example, among the meanings of *see* are INVOLUNTARY-VISUAL-EVENT *and* CONSULT-PROFESSIONAL. The former has much broader selectional constraints than the latter: the

AGENT can be any ANIMAL, and the THEME can be any PHYSICAL-OBJECT. By contrast, the AGENT of CONSULT-PROFESSIONAL must be HUMAN and the THEME must be one of a select inventory of professional roles, including DOCTOR, LAWYER, ACCOUNTANT. By default, the analysis that matches the tighter selectional constraints is preferred: i.e., *I saw my doctor yesterday* will be analyzed as an instance of CONSULT-PROFESSIONAL. However, its score will not be as high as when only one strong candidate analysis is available since the other reading is also possible, as will become clear presently.

2. The meaning of an adjunct invalidates an interpretation (incurs a penalty).

Continuing with our *see* example, the reading with the broader constraints – VOLUNTARY-VISUAL-EVENT – is certainly available: *I saw my doctor in the supermarket yesterday*. In the best case, the context includes an overt clue to override the default reading, such as the adjunct *in the supermarket.* The agent must be on the lookout for such cases, always checking whether explicit indications of LOCATION, INSTRUMENT, and select other case roles align with the currently preferred reading. In this instance, it will consult the ontology for the LOCATION of CONSULT-DOCTOR (the contextually relevant descendant of CONSULT-PROFESSIONAL). Since CONSULT-DOCTOR is not ontologically defined as happening in a SUPERMARKET (it occurs in a HOSPITAL or DOCTORS-OFFICE) the agent penalizes the CONSULT-PROFESSIONAL interpretation. This penalty must be larger than the bonus received by CONSULT-PROFESSIONAL for having tighter semantic constraints. In the end, the INVOLUNTARY-VISUAL-EVENT reading will win, since it does not explicitly contradict any textual or ontological knowledge.

Of course, no single heuristic procedure will resolve all ambiguities in all contexts. For example, when located in a hospital corridor, one can say, *I see my doctor* (visual event) or *I'm here seeing my doctor* (consult professional). This disambiguation decision relies on a combination of physical and linguistic (tense, aspect) clues that require further investigation.

3. Multiple senses have the same selectional constraints (incurs a penalty).

A difficult case occurs when a verb has both physical and metaphorical senses, with both of them participating in the same syntactic structure and imposing the same semantic constraints on case roles. E.g., *He floored me* can mean 'He knocked me to the ground' or 'He surprised me'. If both meanings are recorded in the lexicon, they will participate in the same syntactic structure and impose the same constraints on case roles, differing only in their concept mappings.

The availability of two equally plausible interpretations is a flag for the agent to attempt context-sensitive reasoning. As a first approximation, the agent should attempt to determine which ontological script best aligns with the context, and then search the script for instances of the candidate concepts. For example, a BAR-SCENE script may include a FIST-FIGHTING event whose subevents include HIT, whereas a DISCUSSION script should include an optional CAUSE-SURPRISE event – along with all other concepts referring to causing emotional reactions in others. For domains covered by ontological scripts, such reasoning is feasible; for other domains, the case of paired literal and metaphorical senses will result in residual ambiguity.

4. Unexpected input (incurs a penalty).

Another class of difficult cases involves unexpected input, with "unexpected" being defined in terms of the current state of the static knowledge resources. Typical examples involve lexical lacunae (unrecorded words or word senses) and non-literal language (e.g., novel metaphors and metonymies). Within a given sentence, there can be many combinations of expected and unexpected elements, of which we will present just a sampling for illustration.

Case 1. The verb is unknown and heads a transitive construction whose subject and direct object each have a single sense in the lexicon, and those meanings are quite specific: e.g., PHYSICIAN and MEDICAL-PATIENT. This is a best-case scenario for an unknown verb since the system can search the ontology for EVENTs whose case-roles are filled by the given meanings: e.g., for the input *The physician* [unknown-verbed] *the patient*, the system will find that PHYSICIAN and MEDICAL-PATIENT are the AGENT and THEME, respectively, of MEDICAL-EVENT, as well as many of its subclasses, of course. The system can, therefore, posit that *unknown-verb* is a MEDICAL-EVENT, and create the TMR (MEDICAL-EVENT (AGENT PHYSICIAN) (THEME MEDICAL-PATIENT)). The agent will have moderate confidence that this analysis is correct, yet with the understanding that it is clearly underspecified. Depending on the overall context, this can lead to knowledge discovery by the agent, as by asking its collaborator for more properties of the event or by searching a knowledge base in support of learning by reading (e.g., Nirenburg et al., 2007).

Case 2. The verb is known but is used in a syntactic configuration that has not been recorded. We can draw an example from a recent evaluation exercise that included the following sentence from Arthur Conan Doyle's story, "The Boscombe Valley Mystery": "I found the ash of a cigar, which my special knowledge of tobacco ashes enables me to pronounce as an Indian cigar." Whereas our lexicon included the syntactic expectation *X pronounces Y* Z_{ADJ} *(The doctor pronounced him dead*) this input includes *X pronounces Y* as Z_{NP}.

Syntactic mismatches like these require hypothesizing one or more candidate correlations between syntactic elements and semantic ones. Each correlation must be evaluated semantically. If a correlation fulfills the semantic expectations of the syntactically wrong sense, then it should be accepted and the resulting analysis will receive moderate confidence. In our example, *pronounce* is recorded as meaning a SPEECH-ACT whose THEME is the compositional interpretation of whatever follows. Since virtually anything can be the theme of a speech act, our cigar example will be accommodated.

Case 3. The verb is known and the syntactic configuration jibes with the input, but the listed semantic constraints are not met. This is typified by metonymy: e.g., [In a restaurant] *Give the yellow hat her check.* Multiple senses of *give* involve a BENEFICIARY, but beneficiaries are ontologically constrained to the class of ANIMAL, and no meaning of *hat* is an ANIMAL. This semantic mismatch triggers metonymy processing. The system must search the ontology for the closest relationship between the actual input, HAT, and the expected constraint, ANIMAL. Ideally, it will find HUMAN (AGENT-OF WEAR-CLOTHES (THEME CLOTHING-ARTIFACT)), and CLOTHING-ARTIFACT (SUBCLASS HAT). This creates the conceptual path between HAT and HUMAN, and licenses the hypothesis that *the yellow hat* here can refer to a HUMAN. Although we do not underestimate the challenges in operationalizing metonymy processing, the good news is that many metonymies follow frequent patterns that are readily recorded: e.g., a body part, piece of clothing, or possession can be used to refer to a person, and a part of an object can be used to refer to the object. Metonymic interpretations belonging to recorded classes such as these will receive a much higher confidence rating than interpretations that rely on ontological distance measures.

To reiterate, these were just a few examples of how expected and unexpected input can combine in an utterance, and how unexpected input processing carries various levels of risk and confidence.

5. Wrong lexical root in a multi-word expression (incurs a penalty).

Multi-word expressions, including but not limited to idioms, are recorded by specifying lexical roots associated with syntactic constituents, as shown in our earlier example of *apply pressure to.* Whenever a syntactic element is specified as having a particular root (the direct object must be *pressure* and the head of the PP must be *to*), this is interpreted as a hard constraint. So, an input like *He applied paint to the wall* will not be treated by the cited sense of *apply* since that sense will incur a very large penalty for requiring a lexical item that is not attested in the input.

6. Wrong syntactic features (incurs a penalty).

Syn-struc elements can be associated with syntactic constraints, which are considered hard constraints. For example, in the idiom *kick the bucket,* the direct object must be the word *bucket* in the singular. If the input is *John kicked the buckets*, the number:singular feature constraint is not met and this sense will incur a very large penalty.

7. Adjuncts are accounted for (merits a bonus).

As an aid to automatic disambiguation, the lexicon includes certain adjuncts – i.e., optional arguments. For example, the physical sense of *hit* is transitive and includes an optional PP headed by *with* that indicates the INSTRUMENT of hitting: *She hit the boulder with a stick*. Lexically recording such adjuncts helps the analyzer to disambiguate highly ambiguous prepositions like *with*. If two available lexical senses cover a given input, but one of them includes the adjunct whereas the other would leave the adjunct to be compositionally analyzed, the sense that includes the adjunct receives a bonus. Remember that the adjunct interpretation will only be leveraged if it is both syntactically and semantically appropriate: in the context *She hit the boulder with vigor*, 'vigor' will not be analyzed as the INSTRUMENT because the instrument of a HIT event must be a PHYSCIAL-OBJECT.

To conclude this section, the knowledge recorded in the OntoSem lexicon and ontology offers a rich source of heuristic evidence that contributes to the scoring of lexical disambiguation and semantic dependency determination decisions.

Reference Resolution

We have been developing methods to treat particularly difficult referring expressions, such as elided verb phrases (*She couldn't <u>see the stage</u> but he could **[e]***) and demonstrative pronouns that refer to a proposition (*<u>He is unwrapping the wound</u> but **that** doesn't make sense*) (McShane and Babkin, 2015; McShane, 2015). Our methods do not cover all instances of each type of referring expression since the most difficult cases require sophisticated, domain-specific reasoning. Instead, we are preparing the agent to detect which instances it can treat with high confidence given its current analysis capabilities. Confidence in each reference decision is estimated from the corpus-attested precision of the procedure triggered to resolve the given referring expression.

By way of example, consider one of the resolution strategies for elided verb phrases, which relies on a formalization of the notion "simple parallel context," defined in terms of the output of the Stanford CoreNLP dependency parser. (Manning et al., 2014). For example, *She couldn't <u>see the stage</u> but he could [e]* is simple parallel because the antecedent clause and the ellipsis clause are in a coordinate relationship (parallel), and the antecedent clause includes only one main verb as a candidate antecedent (simple). The system, however, covers many individual cases that differ in the confidence of their prediction. For example, when modality scopes over the proposition in the antecedent clause, it can either be excluded from the resolution (as in the *see the stage* example) or it can be included in the resolution (*He <u>tried to jump rope</u> but I didn't [e]*). This extra decision point introduces room for error. Another potential source of error is a syntactic pruning mechanism that we introduced to increase system recall. This mechanism attempts to convert complex sentences into the simple parallel ones that our approach can treat (McShane et al., 2015). For example, it removes the struck-out portion of the following context and resolves the ellipsis correctly: "'~~We're celebrating the fact that we're living in a time where~~, when we want to *be in the kitchen*, we can [e],' says Tamara Cohen, Ma'yan program director."[2] In order to provide agents with the most accurate estimate of confidence in its reference decisions, we split configurations rather finely in the corpus evaluations that seed those estimates: e.g., simple parallel configurations with no modality and no stripping are separated from simple parallel configurations with modality but no stripping, and so on.

Let us consider one other method of treating ellipsis that involves interesting confidence-related calculations. Some elliptical constructions are recorded in the lexicon, along with corresponding resolution procedures. For example, if the direct object of the verb *start* refers to an ontological OBJECT rather than an EVENT, the associated event has been elided: e.g., *She started a book* means that she started *doing something to/with* a book. There is a special lexical sense of *start* that expects the direct object to be an ontological OBJECT. This sense includes a call to a procedural semantic routine that searches the ontology for the most fitting event that has the textually supplied combination of AGENT and THEME. For the example *She started a book*, the system will search for the event most narrowly specified as having the case-roles AGENT: HUMAN and THEME: BOOK-DOCUMENT. The resolution will often be a set of possible concepts – in this case, READ, WRITE, EDIT and so on. Ideally, further contextual clues will be available to narrow the interpretation. The scoring of the resolution will depend upon whether or not the agent can narrow the interpretation down to 1, and what evidence it brings to bear to do so.

Treatment of Nominal Compounds

We define treating nominal compounds as disambiguating each of the component nominals and establishing the necessary relation – or relations – among them. For example, *cat food* is analyzed as FOOD (THEME-OF INGEST (AGENT CAT)), and *shrimp boat* is analyzed as BOAT (LOCATION-OF CATCH-FISH (THEME SHRIMP)). This depth of analysis stands in contrast to most implemented NN systems, which seek only to select a single relation between uninterpreted nouns.

Our algorithm for treating nominal compounds – detailed in McShane et al., 2014 – involves attempting to analyze each compound using a series of analysis procedures that are ordered according to decreasing confidence.

The most confident analysis obtains when a compound is explicitly recorded in the lexicon, such as *drug_trial*, and *coffee_mug*. Compounds can be lexically recorded due to semantic non-compositionality, frequency of occurrence in text, or relevance for a particular application.

The next resolution strategy involves matching against an inventory of lexically- and/or ontologically-anchored patterns. For example, *FISH* + *fishing* matches a string referring to any kind of FISH followed by the string *fishing*. The analysis is FISHING-EVENT (THEME *the-given-kind-of-fish*). So, *trout fishing* will be analyzed as FISHING-EVENT (THEME TROUT). Similarly, TIME + EVENT matches a string referring to any string analyzed as TIME followed by any string analyzed as an EVENT. So, *Tuesday flight* will be analyzed as FLY-EVENT (TIME TUESDAY). Each pattern in our inventory needs to be tested against a corpus for precision, which will then serve as agent confidence in its analysis of matching compounds.

If a two-element compound does not match any of our recorded patterns, the agent must attempt to find the shortest path between each available interpretation of the first noun and each of the available interpretations of the second noun. This search process offers much lower predictive power than the strategies sketched earlier, and confidence in the results must be accordingly lower. One method of increasing confidence in such analysis involves reference resolution: If one or both of the nouns corefers with an earlier mention, this will help to disambiguate the noun, reducing the "many to many" search space of nominal meanings to "one to many" or, even better "one to one." A quantification method for this heuristic is still to be devised.

To this point, we have assumed that the agent actually knows (i.e., has recorded lexical senses for) both nouns in the compound, and that the needed senses are among the available ones. This, of course, need not be true: the agent might need to learn one or more of the nouns before attempting to analyze the compound. This learning process will be error-prone, further lowering the confidence in the overall interpretation of the compound.

Treatment of Fragments

Sentence fragments are utterances that are not canonical full sentences as defined by normative grammar. They are, however, perfectly natural in ordinary speech. We functionally distinguish three types of sentence fragments:

(1) Those that fulfill a discourse need introduced by the previous utterance: e.g., an answer follows a question, an acknowledgment follows a command. "*What is his blood pressure?" "110 over 80.*"

[2] This example is from Graff and Cieri, 2003.

(2) Those that add additional descriptive information to the previous utterance: e.g., an event mention can be followed by its location, time, or other descriptors. "*The lecture is this afternoon." "At 4:00."*

(3) Those that represent neither of these, such as a sole NP or property serving as an utterance. "[Surgeon] *Scalpel! ... Blood pressure?"*

Our microtheory of processing fragments (McShane et al., 2005) treats fragments of types 1 and 2. It relies on expectations encoded in the TMR. For example, the TMR of a question explicitly includes a property value slot that is waiting to be filled. When a fragment utterance fulfills an expectation like this, processing is quite straightforward and results in a high-confidence analysis.

By contrast, interpreting fragments of type 3 is substantially more difficult, involving script-based reasoning and inferencing about the speaker's goals and plans. If a doctor working on a patient says, "Scalpel!" he or she wants to be handed a scalpel. Although it would be trivial to write a rule saying that anytime people say PHYSICAL-OBJECT! they want to be handed that object, this clearly won't work all the time: *Nuts! Lawyers! My foot!* Constraining the objects only to IMPLEMENTs might help, but would offer less confidence than if the agent could match the meaning of the utterance to a known ontological script. In this case, SCALPEL is an INSTRUMENT-OF SURGERY, a descendant of TREAT-MEDICALLY. The AGENT of SURGERY is, by default a SURGEON, relaxing to a PHYSICIAN or, in rare cases, any HUMAN (these three levels of constraints are recorded in the OntoAgent ontology; the agent is expected to know the social roles of the participants in the conversation); and the PRECONDITION for using any INSTRUMENT to carry out an EVENT is having access to it. The full TMR for *Scalpel!* in this context should be:

REQUEST-ACTION-1
 AGENT HUMAN-1
 THEME TRANSFER-POSSESSION-1
 BENEFICIARY HUMAN-2

TRANSFER-POSSESSION-1
 AGENT HUMAN-2
 THEME SCALPEL-1
 BENEFICIARY HUMAN-1

Procedures for recognizing specific events as parts of scripts are under development. The basic process, however, is clear: the agent will match TMR excerpts from the dialogs against recorded ontological scripts, such that the scripts with the highest number of matches and/or the closest matches will be considered to represent the context of the conversation, thus providing evidence for self-confidence computation. Knowing the context of the conversation can assist not only the interpretation of fragments, but also in lexical disambiguation and reference resolution.

Interpreting Indirect Speech Acts

Indirect speech acts represent a misalignment between the direct and intended meanings of an utterance. We functionally distinguish two classes of indirect speech acts: conventionalized and not conventionalized. The former are recorded in the lexicon and offer high confidence in analysis, whereas the latter require goal-oriented reasoning that is more prone to error. We will consider each class in turn.

Conventionalized indirect speech acts are recorded as phrasals in the lexicon: e.g., the formulations *I'd like to ask you to X, It would be great if you could/would X, I'd really appreciate it if you would X* are all requests for action (ontologically, REQUEST-ACTION THEME: X). A particularly interesting case involves the verb *need*. When used with the subject *you*, it often indicates a request for action (*You need to tighten the bandage*), whereas when used with any other subject it indicates obligative modality scoping over the event (*I need to work faster*). The expressive power of our lexicon readily accommodates such distinctions.

Indirect speech acts that are not conventionalized are challenging, particularly since they are often formulated as statements: e.g., *I'm having trouble doing this* could be a request for help or an instance of talking to oneself during a challenging task. In preparing agents to participate in task-oriented dialogs, we are configuring them to seek, in every dialog turn, a request for information or a request for action. If a given turn already contains a request for action or information, other statements are, by default, assumed to serve as information that the agent must learn: e.g., [*His blood pressure is dangerously low]$_{INFORM}$, [hand me that bandage]$_{REQUEST\text{-}ACTION}$*. If a given turn does not contain any requests, the agent must attempt to match the meaning of the input with its known ontological scripts, and determine if that meaning is closely associated with any actions it can carry out. This takes us deeply into the realm of mindreading and sophisticated agent reasoning (McShane et al., 2013), both of which extend beyond language processing per se and, as such, beyond the scope of this paper.

Final Thoughts

Many methods can be suggested for calculating confidence in intelligent agents, including statistical and probabilistic ones. However, if one's goal is building intelligent agents with human-like capabilities, then such agents must be able to give reasons for their confidence-related assessments. This, in turn, means that an explanatory theoretical model must underlie such judgments.

One way to build an explanatory model – no matter its purview – is a) to propose heuristic features relevant to computing confidence in each individual agent decision that requires a confidence parameter, b) to develop algorithms for computing values of these heuristic features at a specific time during agent functioning, and finally c) to come up with the best ways of combining evidence from the heuristic features relevant to a decision.

In this paper we have presented that part of our explanatory model that deals with decision-making related to language understanding. We have decomposed semantic analysis into a set of core tasks treated by individual microtheories. Each microtheory incorporates a set of heuristic features that are used to estimate confidence in processing individual instances of language phenomena. The unsurprising lesson from pursuing all of these microtheories is: the more lexical and ontological knowledge can be leveraged – and the more specific that knowledge is – the more confident the agent will be in its analysis. For example, an analysis that can leverage a *lexically recorded* phrasal, NN compounding pattern, or indirect speech act will offer higher confidence than an analysis that must rely on more generalized reasoning. This state of affairs serves as a vote for carrying out more high-level knowledge engineering in support of intelligent agents rather than eschewing this necessary work in favor of the currently more popular approach of machine learning, whose results do not nearly reach the level of quality needed to support reliable agents.

All of the aspects of language processing described above either have been implemented or are microtheories at various stages of development within Ontological Semantics. To date, our most advanced agent played the role of a virtual patient in the prototype Maryland Virtual Patient physician training system (Nirenburg et al., 2008; McShane et al. 2013).

Our current and near-future work on agent self-confidence will address issues of quantifying confidence values and combining them to yield a comprehensive confidence judgment for a language input. A further task is to integrate self-confidence assessments in language processing with similar assessments related to the agent's other tasks of perception, reasoning and action.

References

Graff, D., and Cieri, C. 2003. English Gigaword. Linguistic Data Consortium, Philadelphia.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 55-60.

McShane, M. 2015. Expectation-driven treatment of difficult referring expressions. *Proceedings of the Third Annual Conference on Advances in Cognitive Systems*.

McShane, M., and Babkin, P. 2015. Automatic ellipsis resolution: Recovering covert information from text. *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI-15)*.

McShane M., and Nirenburg, S. 2012. A knowledge representation language for natural language processing, simulation and reasoning. *International Journal of Semantic Computing* 6(1).

McShane, M., Nirenburg, S., and Beale, S. 2005. Semantics-based resolution of fragments and underspecified structures. *Traitement Automatique des Langues* 46(1): 163-184.

McShane M., Nirenburg, S. and Jarrell, B. 2013. Modeling decision-making biases. *Biologically-Inspired Cognitive Architectures (BICA) Journal* 3: 39–50.

McShane, M., Beale, S., and Babkin, P. 2014. Nominal compound interpretation by intelligent agents. *Linguistic Issues in Language Technology (LiLT)*, Vol. 10.

McShane, M., Nirenburg, S., and Babkin, P. 2015. Sentence trimming in service of verb phrase ellipsis resolution. *Proceedings of the EuroAsianPacific Joint Conference on Cognitive Science (EAP CogSci 2015)*.

McShane, M., Nirenburg, S., and Beale, S. Forthcoming-a. Language understanding with Ontological Semantics. *Advances in Cognitive Systems*.

McShane, M., Beale, S., and Nirenburg, S. Forthcoming-b. The Ontological Semantic treatment of multi-word expressions. *Lingvisticæ Investigationes.*

Nirenburg, S., and Raskin, V. 2004. *Ontological Semantics*. Cambridge, Mass.: The MIT Press.

Nirenburg, S., McShane, M., and Beale, S. 2008. A simulated physiological/cognitive "double agent". In: Beal J, Bello P, Cassimatis N, Coen M, Winston P, editors., *Papers from the AAAI fall symposium, naturally inspired cognitive architectures*, Washington, D.C., Nov. 7-9. AAAI technical report FS-08-06, Menlo Park, CA: AAAI Press.

Nirenburg, S., Oates, T., and English, J. 2007. Learning by reading by learning to read. *Proceedings of the International Conference on Semantic Computing*. San Jose, August.