7-2019

On true language understanding

Seng-Beng HO

Zhaoxia WANG
Singapore Management University, zxwang@smu.edu.sg

On True Language Understanding

Seng-Beng Ho[1](&) and Zhaoxia Wang[1,2]

[1] Institute for High Performance Computing, A*STAR, Singapore, Singapore
hosengbeng@gmail.com, wangz.ihpc@gmail.com

[2] Nanjing University of Information Science and Technology, Nanjing, China

Published in Artificial Intelligence and Security: 5th International Conference, ICAIS 2019, New York, July 26-28: Proceedings, pp. 87-99. LNCS 11633.
DOI: 10.1007/978-3-030-24265-7_8

Abstract

Despite the relative successes of natural language processing in providing some useful interfaces for users, natural language *understanding* is a much more difficult issue. Natural language processing was one of the main topics of AI for as long as computers were put to the task of generating intelligent behavior, and a number of systems that were created since the inception of AI have also been characterized as being capable of natural language understanding. However, in the existing domain of natural language processing and understanding, a definition and consensus of what it means for a system to "truly" understand language do not exist. For a system to understand an idea, firstly it has to ground the meaning of the concepts in the idea that it manipulates - the concepts that are associated with the words it inputs and outputs. However, there has not been any standardized consensus on what constitutes adequate semantic grounding. This paper presents a spatio-temporal representational method as a basis for a specification of what constitutes adequate semantic grounding, particularly in connection with certain words and concepts related to grounding of physical concepts and mental constructs. This research has critically important implication for learning – true language understanding will usher in an era of learning through language instruction, which is how humans learn, to rapidly accumulate a vast amount of knowledge critical to the propagation of the species and the advancement of its civilization.

Keywords

Grounding of mental constructs, Grounding of physical concepts, Natural language understanding, Semantic grounding, Spatio-temporal representation

1 Introduction

Even though natural language processing (NLP) has been a bourgeoning field in artificial intelligence (AI), and has scored successes in many applications [1–4], natural language understanding (NLU), on the other hand, has not been satisfactorily addressed. With respect to certain NLP applications, such as question-answering, an NLP system is typically able to give reasonably satisfactory answers in many instances. Despite this, from the "understanding" point of view, scientists engaged in NLU research still believe that the system involved does not really "understand." What does it really mean to understand? "To understand" seems to require more than generating a string of language tokens (words) in the output. But what is this extra ability that a language understanding system should have to qualify it as having "truly" understood a language input? What true language understanding is still an unsolved problem and an un-answered question. This paper provides the answer and presents a specification for true language understanding.

2 The Importance of Semantic Grounding

2.1 The Problems with Dictionary Definitions

First, we would like to explore the issue of concept or semantic grounding. Let us begin with a simple concept "Move." Often we refer to the dictionary for "meanings" of words. Here, we would like to show that the dictionary cannot supply the "meaning" for true understanding. The definition of *Move* in Merriam Webster is:

Move: To *go* from one place or position to another

So, for a language processing system to understand the meaning of any concept or word, it has to understand the meaning of the constituents of its definition. Let us next retrieve the definition of a key constituent of the above definition, "Go":

Go: To *move* on a course

Thus, there is circularity in the definitions: *Move* is defined in terms of *Go* and *Go* is defined in terms of *Move*. One has a feeling that the system does not "really understand" other than to rephrase a series of words with another series of words [5].

2.2 The Proposed Spatio-Temporal Representations

What is *really* the meaning of the concept *Move*? We submit that it is a spatio-temporal concept that is better represented in a "pictorial" manner as shown in Fig. 1.

In Fig. 1(a) one can see that there is an axis representing the time dimension and three other axes representing the spatial dimensions. Different from the three spatial dimensions, time is a unique and special dimension sensed by humans. The "blob" is an object A. The object A changes its location over time. Figure 1(b) shows a simplified spatio-temporal representation of the concept *Move* in which the object A changes one unit of elemental location over one unit of time. This is a reduction of a three-dimensional representation (x, y, z) of space (Fig. 1(a)) to a two-dimensional representation (x = 0, y, z) of space for the purpose of simplification without compromising the concepts involved.

In Fig. 1(c) there is a more general representation. The gray "bars" represent "any number of units in between." Therefore, the representation in Fig. 1(c) says that the object A changes *any* amount of space location over *any* amount of time.

Combining the representational schemes of Figs. 1(a), (b) and (c), we define a new space variable l and spatial change Δl, and we leverage the unique and special dimension of time t to represent *Move*. The gray "bars" represent "any number of units in between." Therefore, the representation in Fig. 1(d) says that the object A changes *any* amount of location over any amount of time. This, we submit, is the definition of *Move* in its most grounded and general form.

Of course, pictures such as those in Fig. 1 have to be operated on by some processes to render them fully operational. Therefore, we posit that there are processes that operate on these representations as follows:

(i) RECOGNITION: Firstly, there is the process of recognition. In order to determine whether an instance of *Move* has occurred in the environment with respect to some objects, a system would check if the object has indeed changed location with respect to time, as stipulated by Fig. 1(d)

(ii) ACTION/GENERATION: Secondly, if a system, endowed with the definition of *Move* as stipulated in Fig. 1, is asked to "move the object," it would act to change the object's location over time, as stipulated by Fig. 1(d).

We posit that both (i) and (ii) suffice to demonstrate that the system "truly understands" the concept *Move* as stipulated in the representations of Fig. 1.

Of course, the representation of the concept *Move* need not be pictorial as shown in Fig. 1. One can also use logic language to represent it as follows:

$$\forall \text{Object}, x, y, z, t \quad Location(Object, x, y, z, t) \wedge Location(Object, x+\Delta x, y+\Delta y, z+\Delta z, t+\Delta t) \rightarrow Move(Object) \tag{1}$$

To make the representation simpler and more general, we use l to represent spatial location and Δl to represent spatial location changes, and Eq. (1) can be then represented as:

$$\forall \text{Object}, l, t \quad Location(Object, l, t) \wedge Location(Object, l+\Delta l, t+\Delta t) \rightarrow Move(Object) \tag{2}$$

which states exactly the same thing as the pictorial representations of Fig. 1, which is that if the Object is at location l at time t and then at location $l + \Delta l$ at time $t + \Delta t$, it is deemed to have *Moved,* if $\Delta l \neq 0$. This logic representation, when acted on with some processes, would also be able to render the concept fully operational with respect to the RECOGNITION and ACTION/GENERATION requirements as described above. The critical issue is not whether the fundamental representation that leads to "true" understanding is pictorial or logical. The issue is that space and time are fundamental ad "atomic" with respect to our understanding of the world and only by defining concepts such as *Move*, which is spatio-temporal in nature, in its most fundamental spatio-temporal form, can the true understanding of these concepts be achieved.

2.3 Other Examples of Spatio-Temporal Representations of Physical Concepts

In Fig. 2 we show some other related atomic, ground level concepts represented in spatio-temporal forms.

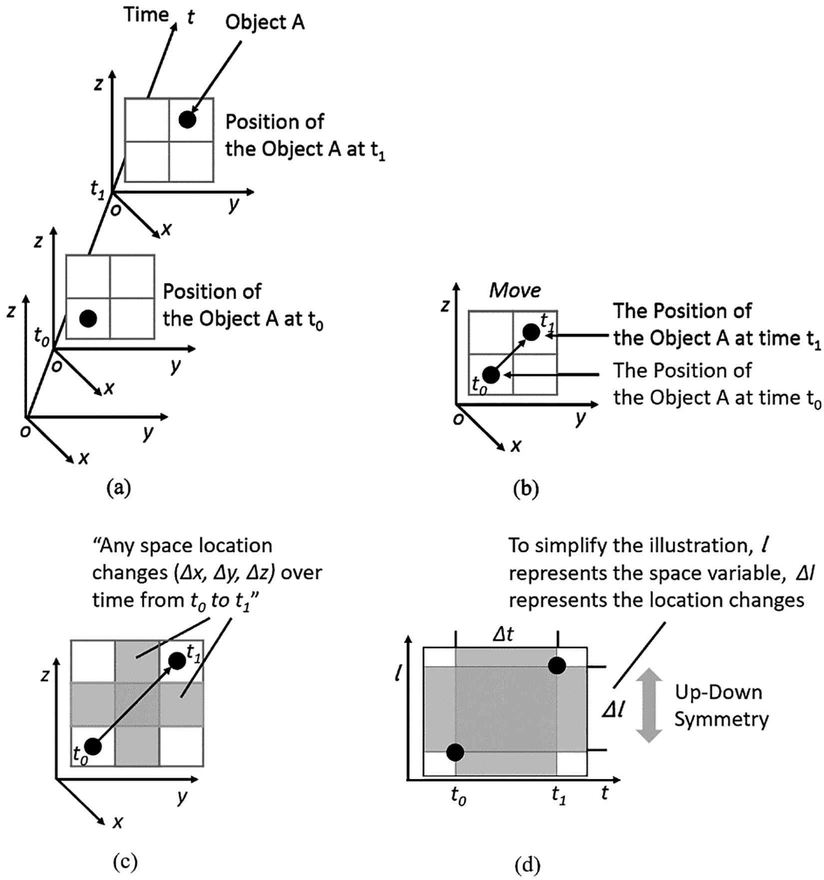

Fig. 1. (a) A spatio-temporal representation of a specific instance of the concept *Move* in which the object A changes one elemental location in both y and z dimensions over time from t_0 to t_1. In order to make the illustration simple, we assume that there is no location change in the x dimension. (b) A simplified representation of the spatio-temporal representation of the concept *Move* in which the Object A changes *one* unit of elemental location in both y and z dimensions over time from t_0 to t_1. (c) A spatio-temporal representation of a general concept *Move* in which the object A changes *any* number of elemental locations in both y and z dimensions over time from t_0 to t_1. (d) After l is leveraged to represent the space dimension and Δl to represent spatial dimensional changes, a spatio-temporal representation of a general concept *Move* in which an Object changes *any* number of elemental locations Δl over *any* amount of time Δt. The "Up-Down Symmetry" indicator specifies that the template has an up-down symmetry – it encodes both the "upward" (i.e., +ve space direction) as well as the "downward" (−ve space direction). Based on [5, 6].

Figure 2(a) shows the concept of "Materialization," in which, at time frame t_1, there is no object in the corresponding location but there is an act of *Materialization*, represented by an "exploding" shape. Then, at time frame t_2, an Object appears. Figure 2(b) shows the concept of "Stay", in which the Object does not change location over time.

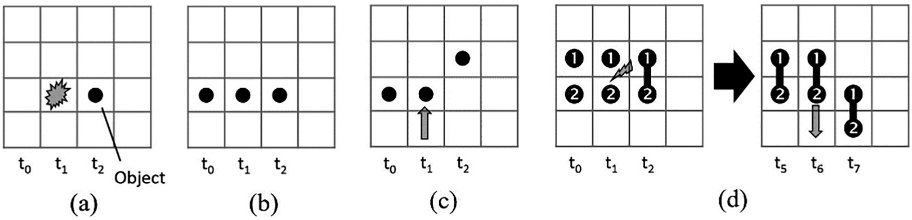

Fig. 2. Spatio-temporal representations of (a) *Materialization*. (b) *Stay*. (c) *Push and Make Move*. (d) *Attach* – the left figure showing the *Attach* action and the right figure showing the consequence – later, if a force is applied to one of them in the direction away from the other, both Objects would move together. The vertical dimension is space and the horizontal one is time. Based on [5, 6].

Figure 2(c) shows the concept of "Push and Make Move" in which at time frame t_1, a force appears and acts on the Object involved, and in the next time frame t_2, the Object changes location by one unit. Figure 2(d) shows the concept of "Attach", in which an *Attach* action is performed at time frame t_1 between two Objects, and in time frame t_2, the two Objects acquire a new property (in a sense a "change of state") which is shown as a bar connecting them. Later, suppose a force is acted on one of them to move it in the direction away from the other, the other will follow. This is the grounded meaning of *Attach* [5, 6].

These are specific instances much like Fig. 1(a). Corresponding generalized versions of these such as that in Fig. 1(b) would allow a system to achieve (i) RECOGNITION and (ii) ACTION/GENERATION as stipulated above. There are other atomic concepts discussed in Ho [5, 6] and we posit that something akin to a set of ground and atomic concepts like this is necessary to achieve semantic grounding for a language system to bring about true language understanding, at least with respect to various physical aspects of the environment. (There are other "mental" aspects such as the various concepts and words referring to the various emotions and sensations experienced by humans that will be discussed below.)

In cognitive linguistics, there has also been some investigations into grounded language understanding. For example, in Fig. 3, the representations of the concepts of "Before" and "After" are demonstrated.

Cognitive linguistics [7] defines *landmark* (lm) as the "focus" of the sentence and *trajector* (tr) as something that is "in reference to the focus." In Fig. 3, it is shown that *Before* and *After* both involve two temporal time frames in a sequence. In the concept of *After*, the earlier event is the *lm* and the later event is the *tr*, and vice versa for *Before*.

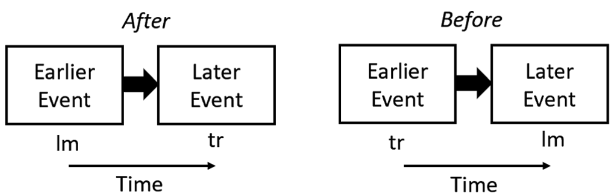

Fig. 3. Cognitive linguistic representation of the concepts of Before and After [7].

2.4 The Issues on the Grounding of Mental Constructs

An important next question would be the issue concerning concepts or words that refer to some mental states such as human sensation or emotion. How would "Saltiness" or "Happiness" be represented? We can certainly represent the consequences of these states, which is *Saltiness* or *Happiness* would lead to certain behaviors of the agents involved (e.g., a sequence of actions to seek more situations that lead to *Happiness* or situations to avoid experiencing *Saltiness*). But what about the concepts themselves? If a human says "I want to *move*," a natural language understanding system can use the above spatio-temporal representation in Fig. 1 to expect a certain spatio-temporal behavior of the human, and hence "truly" understands what he means. However, if he were to say "I am *happy*", and if the system simply responds, "I don't know how it feels to be happy, but I know what you will be doing in the state of happiness," the system can still be a good companion, but the human may then respond, "You don't really understand me."

In Ho [5], it has been demonstrated that it is possible to represent the *changes* of these internal mental states in a system. Consider mental parameters such as *Saltiness* (a kind of sensation) or *Happiness* (a kind of emotion). Their changes can be representation as shown in Fig. 4(a). These changes are in turn describable by the earlier spatio-temporal concepts such as *Move* in the upward direction, which corresponds to *Increase* in the intensity, or *Stay* – no change in intensity.

One can imagine there is a *Saltiness* sensor installed in a robotic system as shown in Fig. 4(b), and the output of the sensor may fluctuate like in Fig. 4(a). Then, the robotic system could output something like, "After consuming salty food, I detect an increase in *Saltiness* in my food receptacle." And if both human and the robotic system agree to the labeling of this particular sensory impression arising from the salty food as "Saltiness", then the robotic system can be said to also "truly" understand the meaning of *Saltiness*. The meaning of *Saltiness* is grounded in the input to the robotic system from the *Saltiness* sensor.

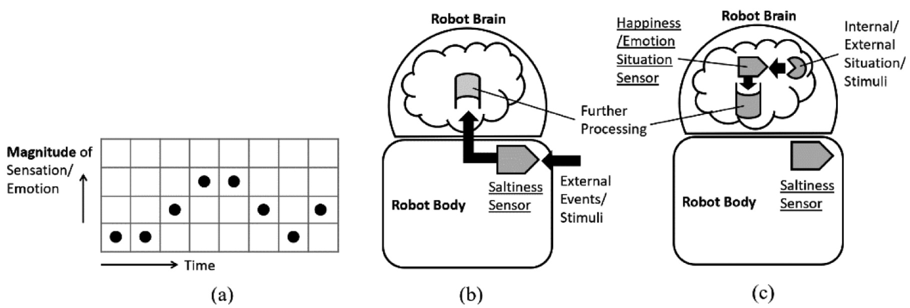

Fig. 4. (a) Changes of some internal mental states such as the *Saltiness* sensation and the *Happiness* emotion. (Increase, decrease, or no change in the intensity of these mental states are shown.) (b) A *Saltiness* sensor for a robot, providing the grounded understanding of the sensation. (c) An internal sensor that senses the emotional situations of *Happiness*, providing the grounded understanding of the emotion.

For the issue of emotions such as *Happiness*, one could imagine that there is an internal mental “Happiness” sensor in a robotic brain, as shown in Fig. 4(c). When certain internal states arises in the brain (e.g., the imagining, recall or sensing of “happy” situations), this sensor detects them and outputs certain signals to signify such a state representing a *Happiness* situation exists. The robot can then be said to understand *Happiness* in the same way that a human does, in parallel with the situation of the understanding of the concept of S*altiness*. The concept of *Happiness* is grounded in the “output” of the internal “sensor” (not from an external sensor like the *Saltiness* sensor) to the rest of the processing system (i.e., some parts of the brain that processes this further). (We ignore the “qualia” problem for now – i.e., the “subjective quality” of the sensation and emotion involved – and instead focus on specifying the *functional* aspects of these sensations and emotions in the systems described in Figs. 4(b) and (c).)

Thus in the case of understanding of certain sensation, the issue cannot be divorced from the presence and availability of certain sensors converting certain external stimuli to internal signal (i.e., if a robot does not have a *Saltiness* sensor, it will never be able to truly understand *Saltiness*, much like a human who is color blind, which means she does not have the corresponding sensors of certain colors, can never truly understand the sensation of these colors). In the case of understanding certain emotions such as *Happiness*, the system must have an “internal sensor” that can sense those corresponding situations (whether currently sensed from the external world or imagined from past experiences), and generate a signal to some other parts of the system that process it like the processing of sensations like *Saltiness*. It is through the detection of this signal that the robot “truly” understands the emotion involved.

2.5 An Example to Demonstrate True Language Understanding

In Ho [5], it has been shown how some instructions can be given to a language understanding system to construct a tool and use it for some purpose. In Fig. 5, we use a simplified 1D (one dimensional) space and 1D time representation to illustrate the process (this should readily generalize to 3D space and 1D time).

In Fig. 5 there is a sequence of actions specified by a stream of linguistic instructions (left side of the figure) and we show the corresponding “understood” actions to be performed (right side of the figure). *We posit that systems like this demonstrate true and deep understanding of the language involved.*

In the figure, the situation begins with an Agent (❶) situated some distance away from an Object (✪) she wishes to retrieve out from a certain Confinement Area. There is a constraint here that the Agent cannot move more than 2 pixel distance from her point of origin (corresponding to the situation that a person may only use her hand to reach out to a certain distance), and the confine is defined as the 3 pixel (1D) space within the current location of the object as shown in the figure. Firstly, it is shown that an elemental object (❷) is being *Materialized* next to the Agent (see Fig. 2 for this operation), effected by the Agent. (In our 3D real world, this could correspond also to the Agent bringing a piece of material to a location for the subsequent construction purposes.) Then, another *Materialization* action is given to materialize another elemental object (❸). Following this, an *Attach* action is generated to attach these two elemental objects together. Subsequently another object (❹) is materialized and

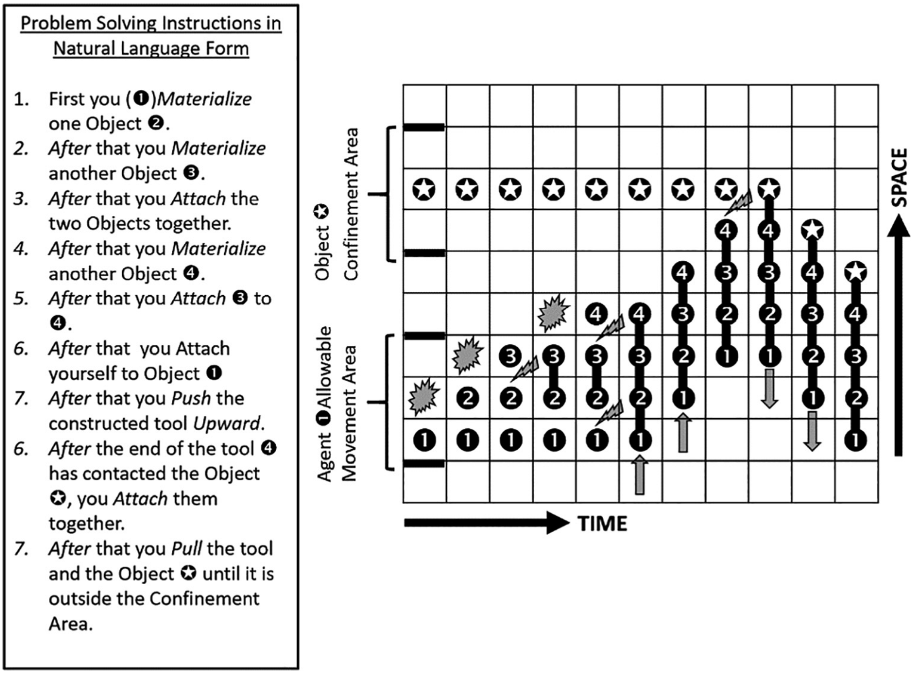

Fig. 5. The understanding of some natural language instructions in a system, which carries out the correct actions accordingly, reflecting its "true" understanding. All the grounded constructs discussed earlier such as *Move*, *Materialize*, *Attach*, *Push*, *After* are brought to bear here. Based on [5].

attached to object ❸. At this point the agent has constructed a long enough tool to reach the desired Object (✪). Under the language instruction, she then *Pushes* the tool "forward" to touch the desired Object, *Attaches* the tool to the Object (in 3D space, the corresponding action could be "grabbing" the object), then *Pulls* the object back toward her initial location. The Object is hence moved out of the Confinement Area.

Through this example, one can see that every part of the language instruction plays a part in instructing the system what to do, and the system "fully understands" in that it carries out the actions accurately. In a sense, *understanding* is *understanding how to act*.

There is another level of the true language understanding system in Fig. 5 that we did not describe, which is the *syntactic* level of understanding – how to convert the English sentences to correctly interpret the overall meaning from the meaning of its parts. We assume that there is a method to do this correctly. What is demonstrated in Fig. 5 is the *semantic* aspect - how the *meaning* of each of the parts of the language instructions on the left side of the figure can be interpreted at the ground level – at the level in which the understanding of the words' meanings lead to specific actions.

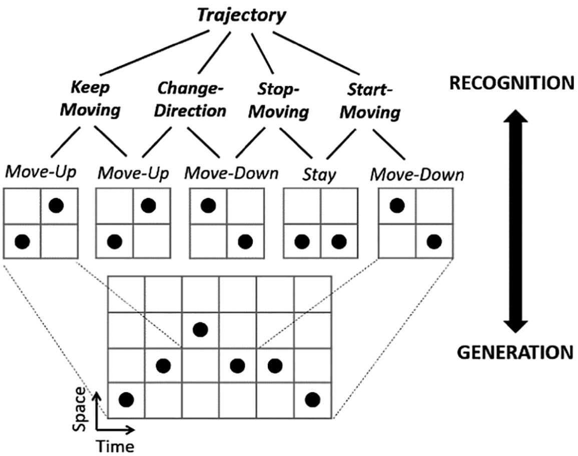

Fig. 6. The higher level concept *Trajectory* defined in terms of lower level and ground level concepts. This is a *specific* instance of the concept of *Trajectory*. It could be suitably generalized like in the case of Fig. 1(b). RECOGNITION and GENERATION of the concept are quite straight forward. Based on [5, 6].

2.6 Higher Level Concepts Are Grounded Through Lower Ones

Concepts such as *Move* are at the ground level in that they have *direct* spatio-temporal correspondences. There are concepts that are at much higher levels that are built upon these concepts. For example, "Change-Direction" can be defined in terms of *Move-Up* and *Move-Down* as shown in Fig. 6. At a higher level, there is the concept of "Trajectory." Figure 6 is a *specific* instance of *Trajectory* [5, 6]. Something akin to the general version of *Move* in Fig. 1(d) can be concocted for *Trajectory* as well and it can be defined as a sequence of "*any number of Moves, any number of Changes of Directions, any number of steps of Stay, in any combination*." But its meaning is grounded in the ground level constructs of *Move*, *Stay*, etc. Thus, higher, more abstract levels of concepts are grounded through intermediate and ground level concepts.

Figure 6 also shows the construct can be used to RECOGNIZE as well as GENERATE the concept of *Trajectory* in a very straight forward manner.

The basic dictionary definition of Trajectory (Merriam-Webster) is:

Trajectory: The curve that a body (such as a planet or comet in its orbit or a rocket) describes in space.

Compared to the representation in Fig. 6, the representation in Fig. 6 is grounded in the sense that it satisfies the RECOGNTIION and ACTION criteria above in a very straight forward manner, and allows a system to directly operate on it.

3 Earlier Natural Language Understanding Work in AI

3.1 Winograd's SHRDLU

In earlier AI work such as that of Winograd [8], true language understanding such as that posited above has actually been achieved to some extent. Figure 7 shows what his system, called SHRDLU, is able to do. Given some commands in language form, the system is able to interpret the commands and carry out actions in the real world to demonstrate that those commands are understood. Words such as "Pick-up," "big," "black," "box" are defined and understood at the ground level. The knowledge is hand-coded in procedural form.

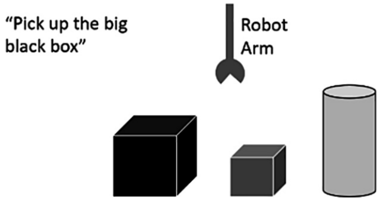

Fig. 7. SHRDLU – an earlier natural language understanding effort. Based on [8].

3.2 Schank's Script Representation

Another work in earlier AI by Roger Schank and Robert Abelson [9] used "scripts" to encode deep understanding. Figure 8 is an example of a "Restaurant Script" in which all the activities that take place in a restaurant together with the corresponding goals and intentions of the participants are encoded (Fig. 8 is a vastly stripped-down version of the script. Please refer to [9] For the full detailed version).

Restaurant Script

START: Agent is *Hungry*. Agent has money
END: Agent is not *Hungry*. Agent has less money

1. SCENE 1 **ENTERING**
   1) Agent looks around for **Table**
   2) Agent decides where to sit
   3) Agent ***Moves*** to **Chair**
   4) Agent ***Sits***
2. SCENE 2: **ORDERING**
   1) ... (DETAILED ACTION SEQUENCES)
3. SCENE 3: **EATING**
   1) ... (DETAILED ACTION SEQUENCES)

Fig. 8. Schank and Abelson's [9] Restaurant Script (a stripped-down version of a figure in [9]).

With this deep representation, their question answering system can respond to statements such as “I went to the restaurant yesterday. I didn’t leave a tip.” The expected human-like response would be “Oh, was the service bad?” Now, a typical dictionary definition of the concept of a restaurant is “a business establishment where meals or refreshments may be purchased” (Merriam Webster). With this kind of shallow, ungrounded definition and representation, the above human-like response would not be possible.

3.3 Relationships Between the Earlier Effort and Ours

The difference between our proposed true language understanding representations as demonstrated in Figs. 1, 2, 3, 4, 5 and 6 and that of Winograd and Schank is: our is at an even deeper ground or atomic level representation. For example, the “block” in Winograd’s SHRDLU system or the Table in Schank’s Restaurant Script have further detailed structures that are not captured in their representations. But in spirit, our system is similar to theirs.

The research in language understanding such as that represented by the work of Winograd and Schank fizzled out because their systems were not scalable. There were no computer vision and machine learning at that time (the 1970’s) and the knowledge involved was hand-coded. If an AI system is endowed with computer vision and learning capabilities, deep knowledge structures such as the Restaurant Script can be learned by bringing the system (say, a “child” robot) to a restaurant and let it observe and learn the activities inside. As the robot moves about in an environment, it will be able to pick up scripts of all kinds of activities. These scripts will then form the knowledge basis for deep and true language understanding.

Pei et al. and Si et al. [10, 11] have demonstrated just such a computer vision capability. Through video observation of the activities in a room carried out by various human agents, their system is able to construct a causal spatio-temporal AND-OR graph that encodes all the possible observed sequences of long range human behavior in the scene. This is akin to the scripts of Schank and Abelson, except that this also represents both the learning and encoding of grounded information about the real world. This will set the stage for the acquisition of knowledge for true language understanding.

4 Summary and Conclusion

In this paper we first showed that dictionary-like definitions, i.e., definitions of words in terms of other words, are not sufficient for a language processing system to achieve true language understanding because these definitions are ultimately circular. To achieve true understanding of various concepts and words, these concepts and words must be grounded, either directly or indirectly, through other intermediate concepts to some ground level representations that are directly tied to the physical or mental constructs to which they refer. We presented the idea that for physical concepts, there is a set of spatio-temporal descriptions that can suffice to provide the ground level representations (e.g., *Move*, *Materialize*, etc.). We posited that there is probably a limited, small set of these “atomic” basic concepts that are sufficient to ground all the concepts related to physical concepts in the human lexicon, and this is consistent with previous work.

For mental constructs such as sensation and emotion, we posited that an intelligent system needs to have the corresponding external and internal sensors akin to those humans possess to be able to truly understand the corresponding concepts or words in a natural language discourse.

We also specified the conditions of being able to use these concepts both to recognize instances of them in the real world and to act on the real world to be the conditions for “true” understanding. Thus understanding, in a sense, is understanding of how to recognize and act. We reviewed two former natural language understanding systems from the early days of AI, namely that of Winograd and Schank and Abelson, to demonstrate that similar ideas have been propounded before but were forgotten in AI because those systems were not scalable. We also mentioned that armed with the new tools of computer vision and machine learning, systems like these may become scalable and therefore we should revisit this issue of natural language understanding.

Future research would focus on (i) further ascertaining a sufficient set of grounded concepts for most if not all concepts/words in the human mental lexicon to be grounded on; (ii) elucidating on how other complex concepts can be grounded accordingly; (iii) demonstrating an AI or robotic system that can really benefit from true language understanding – either in carrying out complex instructions by humans correctly, or through language, learning a vast amount of knowledge necessary for its functioning, much like how the vast majority of a human’s knowledge is learned.

This will usher in an era of machine learning akin to that of human learning, in which most complex knowledge is learned rapidly through language instruction. This kind of machine learning will be set totally apart from the slow and restricted kinds of machine learning today – the supervised and unsupervised learning methods applied to pattern recognition, or the reinforcement learning method applied to action sequence learning.

References

1. Ferrucci, D., et al.: Building Watson: an overview of the DeepQA project. AI Mag. **31**(3), 59–79 (2010)
2. Ganegedara, T.: Natural Language Processing with TensorFlow: Teach Language to Machines Using Python’s Deep Learning Library. Packt Publishing, Birmingham (2018)
3. Manning, C.D., Schutze, H.: Foundations of Statistical Natural Language Processing. MIT Press, Cambridge (1999)
4. Wang, Z., Chong, C.S., Lan, L., Yang, Y., Ho, S.-B., Tong, J.C.: Fine-grained sentiment analysis of social media with emotion sensing. In: IEEE Future Technologies Conference 2016 (FTC 2016), San Francisco, United States, 6–7 December 2016
5. Ho, S.-B.: Principles of Noology: Toward a Theory and Science of Intelligence. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-32113-4
6. Ho, S.-B.: The atoms of cognition: a theory of ground epistemics. In: Proceedings of the 34th Annual Meeting of the Cognitive Science Society, pp. 1685–1690. Cognitive Science Society, Austin (2012)

7. Langacker, R.W.: Foundation of Cognitive Grammar, vol. I. Stanford University Press, Stanford (1987)
8. Winograd, T.: A Procedural Model of Language Understanding. In: Schank, R., Colby, K.M. (eds.) Computer Models of Thought and Languag. W. H. Freeman & Company, San Francisco (1973)
9. Schank, R., Abelson, R.: Scripts, Plans, Goals, and Understanding. Lawrence Erlbaum Associates, Hillsdale (1977)
10. Pei, M., Jia, Y., Zhu, S.-C.: Parsing video events with goal inference and intent prediction. In: International Conference on Computer Vision. IEEE, New Jersey (2011)
11. Si, Z., Pei, M., Yao, B., Zhu, S.-C.: Unsupervised Learning of AND-OR grammar and semantics from video. In: International Conference on Computer Vision. IEEE, New Jersey (2011)