Commonsense inference in human-robot communication

Aliaksandr Huminski
Institute of High Performance Computing,
Singapore
huminskia@ihpc.a-star.edu.sg

Ng Yan Bin
A*STAR AI, Singapore
Singapore
ng_yan_bin@scei.a-star.edu.sg

Kenneth Kwok
Institute of High Performance Computing,
Singapore
kenkwok@ihpc.a-star.edu.sg

Francis Bond
Nanyang Technological University,
Singapore
fcbond@ntu.edu.sg

Abstract

Natural language communication between machines and humans are still constrained. The article addresses a gap in natural language understanding about actions, specifically that of understanding commands. We propose a new method for commonsense inference (grounding) of high-level natural language commands into specific action commands for further execution by a robotic system. The method allows to build a knowledge base that consists of a large set of commonsense inferences. The preliminary results have been presented.

1 Introduction

There is a significant progress in movement from early natural language understanding computer programs like SHRDLU (Winograd, 1972) with its deterministic actions in the virtual world to modern cognitive robots operating in the physical world and mapping language to actions. Artificial agents enter our lives and the end users of such systems are not technical experts. The only way for them to communicate with AI is to use natural language. For example, humans can give a natural language command expecting a follow-up action by the agent.

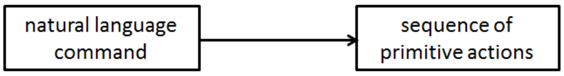


Figure 1: Transformation of high-level command for an agent.

Nowadays in robotics, in order to execute a natural language command which is considered as a high-level instruction, an agent needs to transform it to a sequence of lower-level primitive actions (Figure 1.). For example, the industrial arm SCHUNK has three primitives: *open-gripper*, *close-gripper*, *move-to* and for this agent any high-level command should be transformed into a sequence of these 3 actions to be performed (Kress-Gazit et al., 2008). For smarter agents with more primitives, complicated commands like *fill up the cup with water* can be executed by transformation into a long sequence of the lower-level actions: *pick up the cup, move to your left, put the cup under the faucet, turn on the faucet, turn off the faucet*, etc. In other words, natural language command decomposition is a necessary step for an agent to be capable of executing.

To make such transformations possible, previous works (Misra et al., 2015; She and Chai, 2016) explicitly model verbs with predicates describing the resulting states of actions. Their empirical evaluations have demonstrated how incorporating result states into verb representations can link language with underlying planning modules for robotic systems (Gao et al., 2016). Recent investigations use reinforcement learning to transform language commands into primitive actions (Misra et al., 2017) or representation of actions (Arumugam et al., 2017).

The current studies in human-robot communication (She and Chai, 2017; Chai et al., 2018) show that natural language understanding of commands is difficult for machines because commands in human-human communications are usually expressed through a desired change of state.

2 Problem Statement

As Rappaport Hovav and Levin (2010) pointed out, any action can be expressed in two different ways. Firstly, there are manner verbs that describe how actions are carried out – i.e. manners of doing: *hit, stab, scrub, sweep, wipe, yell*, etc. Secondly, there are verbs that describe results of an action or a change of state: *break, clean, crush, destroy, shatter*, etc.

Further we will use a term "action verb" as a

Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing, pages 104–112
Hongkong, China, November 3, 2019. ©2019 Association for Computational Linguistics

synonym for a manner verb and a term "result verb" as a synonym for a verb that describes a result of an action or a change of state.

For commands in human-human communication, people mostly use result verbs. We say *open the door*, not *push the door*; *clean the table*, not *wipe the table*.

It should be underlined that result verbs don't express any concrete action. For instance, the command *open the door* represents a particular kind of change of state in an entity but it is silent about **how** this change comes about. The verb *clean* doesn't indicate whether it was done by sweeping, wiping, washing or sucking; the same way the verb *kill* does not indicate how a killing was done[1].

On the contrary, the action verbs in the commands *pull the door, push the door, kick the door*, etc. represent different kinds of action necessary to implement the change of state *open the door*.

The obvious question arises: if a command is expressed through a desired change of state, how humans know what actions to do? The point is that humans derive the information about the concrete actions related to the desired change from shared background knowledge about the world. There is no need to explicitly represent it in human communication. It is commonsense knowledge that enables us to understand each other (Clark, 1996; Tomasello, 2008) and to know how to open the door or how to clean the table (see Figure 2.).

AI systems, even new generations of cognitive agents, have significantly less knowledge about the world and are not able to ground result-verb commands into action-verb commands. A command with a result verb does not give AI any information on *what actions* should be performed to achieve the disable change of state. As a result of that, commands to robots are directly linked to primitive actions implemented by a robot without the intermediate step of identifying them with action verbs (see Figure 1.).

The straightforward approach "command → primitive actions" fails to achieve two significant points.

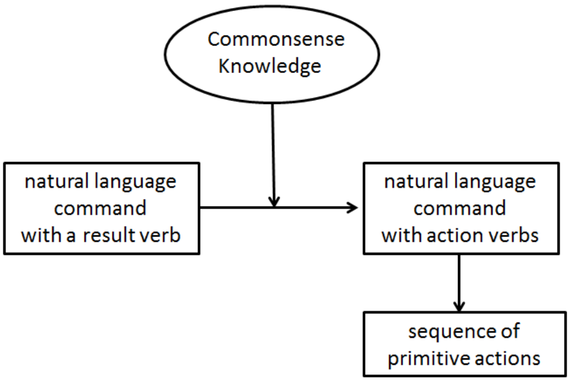

Commonsense Knowledge
natural language command with a result verb
natural language command with action verbs
sequence of primitive actions

Figure 2: Transformation of high-level command for a human.

First, a result verb being applied to the same object can be executed by different action-verb commands. For instance, the command with the result verb *fill up (the cup with water)* can be executed by the action verb *pour (water into the cup)* or by the action verb *scoop (water from the bucket)*.

Second, a result verb being applied to different objects assumes different action verbs. For instance, the following commands with the same result verb *open* require different action verbs to be executed: *open the door; open the book; open the refrigerator; open the can; open the envelope*, etc. Even for the similar commands *open the door* and *open the refrigerator* there is a difference that must be noted: the last command cannot be implemented by pushing.

The general problem of overcoming the gap in human-robot natural language understanding being applied to the high-level natural language commands can be formulated the following way. How can AI systems transform high-level natural language commands with result verbs into commands with action verbs[2]?

3 Related Work

Although commonsense inference between action verbs and result verbs has been described in linguistic studies (Rappaport Hovav and Levin, 2010), there is still a lack of detailed account of potential causality that could be denoted by an action verb (Gao et al., 2016).

From the AI domain, there were investigations

[1] The separation of verbs on action verbs and result verbs got further elaboration in cognitive science where an event representation is considered to be based on 2-vector structure model: a force vector representing the cause of a change and a result vector representing a change in object properties (Gardenfors, 2017; Gardenfors and Warglien, 2012; Warglien et al., 2012). It is argued that this framework gives a cognitive explanation for manner verbs as force vectors and for result verbs as result vectors.

[2] In the article we do not consider the follow-up step in the transformation of action verbs into action primitives for further execution by AI agent. This kind of transformation depends on the type of the agent.

devoted to learning the physics of the world from videos (Fire and Zhu, 2016) and simulations (Wu et al., 2017). However, except for a few works that explored the physical properties of verbs (Forbes and Choi, 2017; Zellers and Choi, 2017), how verbs and their corresponding actions affect the state of the physical world is still largely underexplored.

Well-known knowledge bases like Freebase, YAGO or DBPedia, even being automatically populated by modern NLP methods, do not contain commonsense inferences we are going to create.

Crowd-sourcing resources such as ConceptNet have an incomplete coverage, which is its main drawback. A human knowledge engineer may not list all possible events related to a particular action verb or a result verb. For example, the inference *scrub* → *clean* might be listed while others such as *mop* → *clean*, *suck* → *clean*, or *sweep* → *clean* might be missed.

Existing linguistic resources such as Propbank, FrameNet or VerbNet provide important information about verb classification, its arguments and semantic roles, but they do not distinguish action verbs and result verbs. For instance, in the largest domain-independent computational verb lexicon VerbNet (Kipper Schuler, 2005), that provides semantic role representation for 6394 verbs (version 3.2b), the action verb *hit* and the result verb *break* have the same structure: [Agent, Instrument, Patient, Result]. Even if the semantic representation for a verb may indicate that a change of state is involved, it does not provide the specifics associated with the verb's meaning (e.g., to what attribute of its patient the changes might occur) (Gao et al., 2016).

WordNet, manually created by professional linguists, to the best of our knowledge, is the only linguistic resource that partly provides information about causal links between action verbs and result verbs. As we will indicate below, these links overlap with the hypernym-hyponym relations in WordNet.

Finally, the broad-coverage resource VerbOcean (Chklovski and Patel, 2004) set a semantic relation "enablement" between verbs using the following 4 patterns: "Xed * by Ying the"; "Xed * by Ying or"; "to X * by Ying the" and "to X * by Ying or", where "X" and "Y" are verbs; (*) matches any single word. The patterns are similar to the one we are going to use. The only significant difference is that all of them do not include a noun after a verb "X". As it was mentioned in the section 3 (2nd point), a result verb being applied to different objects assumes different action verbs.

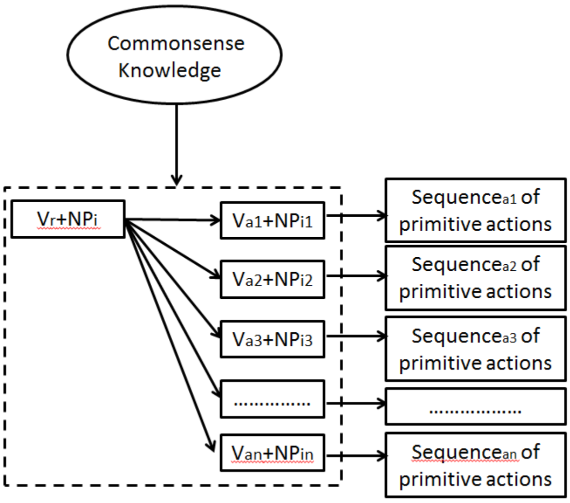

Commonsense Knowledge
V_r+NP_i
V_{a1}+NP_{i1}
V_{a2}+NP_{i2}
V_{a3}+NP_{i3}
....
V_{an}+NP_{in}
$Sequence_{a1}$ of primitive actions
$Sequence_{a2}$ of primitive actions
$Sequence_{a3}$ of primitive actions
....
$Sequence_{an}$ of primitive actions

Figure 3: Transformation of high-level command.

4 Proposed Approach

We consider the transformation formulated in section 3 as a process of grounding where a high-level command representing a desired change of state is grounded to an action(s) command.

The following two assumptions will be made to formalize the process of grounding.

1) The commands in human-robot interactions can occur in various forms and patterns. Some of them can be rather complicated. Our work addresses the simplest case where a command is represented by the structure V+NP, where V is a verb, NP is a noun phrase.

2) The grounding of a result-verb command into an action-verb command is represented as: V_r+NP_1+by+V_a+NP_2, where V_r is a result verb; V_a is an action verb[3].

Since a result verb being applied to the same object can be executed by different action-verb commands, the schema on the Figure 2. will be unfolded as one-to-multi relations between a result-verb command and an action-verb command (see Figure 3.).

The key point here is how to extract one-to-multi relations. In reality, these relations are commonsense inferences that allow humans easily to

[3] NP_1 can be the same or different from NP_2.. Compare: *open the door by pulling the door* and *open the door by pushing the button*

transform result-verb commands into action-verb commands. These commonsense inferences are so obvious and so well-known to everybody that are very rarely expressed anywhere in a written form. It makes it hard to find and extract from any source of information. As a consequence of that, we cannot apply deep learning techniques for extraction of above-mentioned one-to-multi relations. Deep learning has proved incredibly powerful and effective for many practical tasks from perceptual classification to self-driving cars. But we have to acknowledge the data-hungry nature of systems based on deep learning. The side-effect of that is a long tail of low-frequency data that cannot be treated the same way. Our research deals with such data.

The method suggested for one-to-multi relations extraction is based on 3 non-related approaches and includes three steps accordingly.

1. Getting 2 sets of verbs: a set of result verbs $\{V_r\}$ and a set of action verbs $\{V_a\}$;

2. Getting a set of the most frequent pairs $\{V_r+NP\}$;

3. Getting a set of commonsense inferences $\{V_r+NP_1+by+V_a+NP_2\}$.

In the first step, result verbs (V_r) and action verbs (V_a) are separated. The separation is based on analysis of Wordnet; this is a domain-independent step that aims to cover generally result and action verbs representing the physical world. In the second step, the set of the most frequent pairs $\{V_r+NP\}$ are extracted using the N-gram approach to form result-verb commands: *clean the floor, cool the beer*, etc. In the third step, we use a search engine to check all around the web if there is a commonsense inference between an action-verb command and a result-verb command (*open the door by pressing the button*). If a commonsense inference exists in the web it is considered as being validated and added to the set.

4.1 Step #1: Getting Two Sets of Verbs

The output of the step #1 is two sets of verbs: a set of action verbs $\{V_a\}$ and a set of result verbs $\{V_r\}$. The separation is based on the analysis of the entire set of verbs through Princeton WordNet (WN) (Fellbaum, 1998) which is widely used in a variety of tasks related to extraction of semantic relations. The verb part of WN contains 11529 unique verbs (version WN 3.0)[4]. They are organized in verb synsets ordered mainly by troponym-hypernym hierarchical relations (Fellbaum and Miller, 1990). According to the definitions, a hypernym is a verb with a more generalized meaning, while a troponym replaces the hypernym by indicating a manner of doing something. The closer a verb is to the bottom of a verb tree, the more specific the manners that are expressed by troponyms: *communicate-talk-whisper*[5].

Meanwhile, action verbs are hidden in the WN verb structure since troponyms are not always action verbs. In some troponym-hypernym relations the verbs are in fact action verbs like in {kill}-{drown}. However, there are no explicit ways to extract them yet.

The idea is that action verbs can be extracted from WN if at least one of four conditions, applied to a verb is valid[6] :

1. A verb in WN is an action verb if its gloss contains the following template: "V + by [...]ing", where V=hypernym. Example: {sweep} (*clean by sweeping*);

2. A verb in WN is an action verb if its gloss contains the following template: "V + with + [concrete object]", where V=hypernym. Example: {brush} (*clean with a brush*). Restriction on the concrete object is made to avoid cases like *with success (pleasure, preparation*, etc).

3. A verb in WN is an action verb if it represents movement in any direction: *lift, turn, descend*, etc.

4. A verb in WN is an action verb if its hypernym is an action verb. In other words, once a verb is an action verb, all branches located below consist of action verbs as well, regardless of their glosses.

The procedure of using conditions 1-4 goes from all top verbs to the bottom verbs. For ex-

[4] https://wordnet.princeton.edu/documentation/wnstats7wn. The following paper (McCrae et al., 2019) outlines a roadmap for adding new entries to WordNet, so the number of verbs is not fixed, but increasing over time.

[5] Note that these are defined on verb-senses, not verbs. For example, the verb *see* "perceive: I see the picture" will behave differently from the verb *see* "understand: I see the problem".

[6] These 4 conditions elaborate the approach developed in (Huminski and Zhang, 2018a,b)

ample, we start from the top synset {change, alter, modify} (gloss: *cause to change; make different; cause a transformation*). It doesn't satisfy the 1st or the 2nd condition, so we go down on 1 level and examine one of its troponyms: {clean, make clean} (*make clean by removing dirt, filth, or unwanted substances from*). It is still not an action verb synset: in the pattern from the 1st condition – "V + by [...]ing" – the verb *make clean* is not a hypernym. On the next level there are synsets with glosses that satisfy either the 1st or the 2nd condition:

- {sweep} (*clean by sweeping*);
- {brush} (*clean with a brush*);
- {steam, steam clean} (*clean by means of steaming*).

So, the verbs *sweep, brush, steam, steam clean* are action verbs. Applying the 3rd condition on them, one can state that all synsets located below these 3 synsets (if any) are action verb synsets. The framework is the basis of the procedure for action extraction.

We implemented the procedure following the conditions 1.-4. and got the following results:

1. 191 verb synsets have been extracted by matching the template "V + by [...]ing";
2. 329 verb synsets have been extracted by matching the template "V + with + (a/an)? + ..." ;
3. 1408 verb synsets have been extracted from the motion lexicographer file;
4. a total of 3063 verb synsets have been extracted as a total number of action verbs including all the verb synsets that are located under the hypernyms as action verbs; 3063 extracted verb synsets contain 3294 unique action verbs.

All other verbs are potentially result verbs. Also some restrictions need to be applied to consider only the result and action verbs that are represented in the physical world and necessary for robot actions.

We will evaluate the results intrinsically (a linguist will judge the validity), and extrinsically, i.e. for English verbs also found in Levin's *English Word Classes and Alternations* (1992) we will compare our results to her classes. For example, class 10.3 "clear" verbs (*clean, clear, drain, empty*) are result verbs while 10.4.1 "wipe" verbs (*bail, buff, dab, distill, dust, erase, expunge, flush, leach, lick ..*) are action verbs.

4.2 Step #2: Getting Set of Pairs {$\mathbf{V_r}$+NP}

The output of the step #2 is a set of the most frequent (commonly used) pairs {V_r+NP}. The purpose of this step is based on the observation that a result verb being applied to different objects assumes different action-verb commands.

To generate the set {V_r+NP} we use N-grams (which are a contiguous sequence of n items from a given text) extracted from the largest publicly-available, genre-balanced corpus of English: the Corpus of Contemporary American English[7] with about 430 million words in size. With this N-grams data (2, 3, 4, 5-word sequences, with their frequency), the subset of N-grams are extracted where the 1st word is a result verb in any grammatical form. A threshold was set for the frequency. For example, for the result verb *open* we extracted all 3-grams that look like the following (with frequency at the beginning):

3459 opened the door
2611 open the door
....
201 open the window
169 opened the window
......
130 opened the box
89 open the box
etc.

If the data from N-grams is insufficient we use larger, noisier corpora such as the common crawl[8].

4.3 Step #3: Getting a Set of Commonsense Inferences

The output of the final step #3 is a set of commonsense inferences between an action-verb command and a result-verb command validated by a search engine from the web. The search engine is used to check (validate) if a commonsense inference exists in the web. Each commonsense inference for the checking has a structure V_r+NP_1+by+V_a+NP_2 (*open the door by pressing the button*).

The procedure is the following:

[7] https://www.ngrams.info/

[8] https://commoncrawl.org/

1. make a cartesian multiplication of pairs {V_r+NP} and action verbs {V_a}: {(V_r+NP), V_a};

2. create a sequence for each element from 1.: V_r+NP+by+V_a (*fill the cup by pouring*);

3. run the sequence from 2. on the search engine looking for the sequence V_r+NP_1+by+V_a+NP_2(concrete object) in the web. Estimate the frequency (or getting no result).

4. If we do not find sufficient action-verb templates V_a+NP_2(concrete object), we will use the learned combinations to learn new templates, extending the approach (Snow et al., 2006) to learning wordnet relations.

All validated commonsense inferences will be added to the set with frequencies and stored.

5 Implementation and Preliminary Results

The flowchart (Fig. 4) shows the general approach of causal relations extraction from text. Three modules on the bottom in grey color represent three steps from section 5. The details of the approach are given below.

Raw data. WordNet is used as raw data.

Algorithm of separation. For getting preliminary results, commonly used result verbs and action verbs were taken from the linguistic literature. We extracted 12 result verbs and 50 action verbs.

Result verbs: *break, clean, clear, close, raise, cut, fill, heat, kill, lift, open, remove.*

Action verbs: *blow, brush, chip, chop, clip, comb, compress, drown, flap, grab, grasp, grind, grip, hack, hammer, hit, kick, knead, lever, mow, pound, pour, press, pull, push, rinse, rub, saw, scoop, scrape, scratch, scribble, scrub, shake, shave, shoot, shovel, slap, slash, smear, soap, splash, sponge, squeeze, stab, steam, sweep, touch, wash, wipe.*

N-gram approach. For each of 12 result verbs, we extracted five 3-grams V_r+NP. Each 3-gram contains the most frequent noun phrase with the corresponding verb. Totally 60 3-grams were extracted (see Table 1 for details).

Web-search. Cartesian multiplication of 60 3-grams and 50 action verbs produces 3000 combinations "V_r+NP by V_a". We use search engine Bing for running the template "V_r+NP by V_a...". Accordingly, 3000 searches were made. The results were taken and analyzed from the first 10 web pages that appeared. We were looking for the results corresponding the template "V_r+NP by V_a+NP/Pronoun".

Results. As a result we got 497 causal relations. Sample of 20 extracted causal relations is given in Table 2.

Examples of causal relations for the 3-gram "open the window" is given in Table 3.

6 Evaluation

The evaluation was based on a sample of 100 causal relations randomly taken from extracted 497 ones.

Due to the restrictions applied on event and causal relation between events we can not evaluate the recall of the extraction.

The precision (validity) of extracted causal relations were evaluated by five human judges. They were given instructions to rate the causal relations by marking each relation with a number from 1 (very bad) to 5 (very good). Examples of invalid (*break the ice by seeing it*) and valid (*opened the box by pulling on the handle*) causal extractions were provided.

6.1 Simple Average

After 5 judges put their marks, the simple average was calculated by dividing the sum of all marks by 500. We got 3.1.

6.2 Extraction of valid causal relations

We calculated the average between judges for each causal relation and extracted 62 causal relations (among 100 randomly taken) with average score more or equal 3.

6.3 Analysis of invalid causal relations

38 causal relations with the average score lower than 3 were preliminarily analyzed for detecting the reasons. We found the following:

a) bad parsing or bad POS tagging (*kill the bacteria by pouring a half cup; fill the hole by pushing thousands; open the window by grabbing the opening*);

b) unusual causal relations that require a context: *heat the oil by pressing the palms; cut the engine by pulling both paddles.*

c) meaningless causal relations: *break the ice by seeing it; killing each other by slashing the rate;*

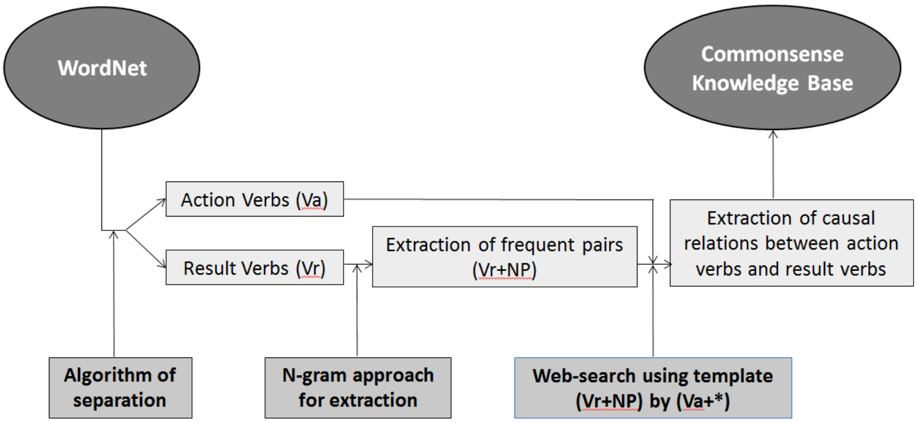
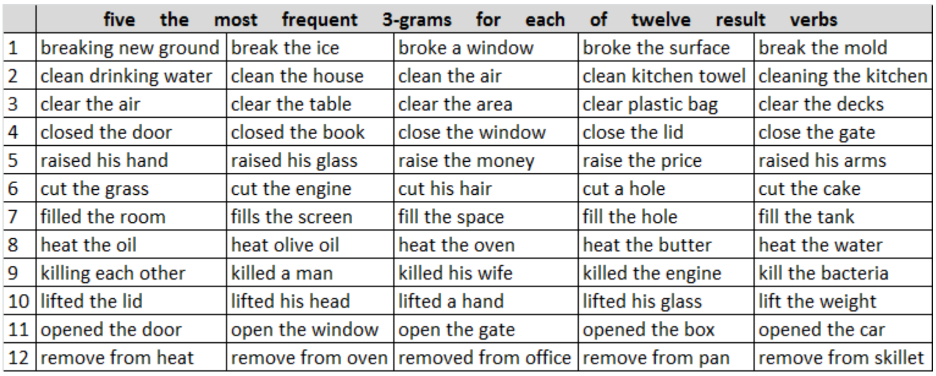
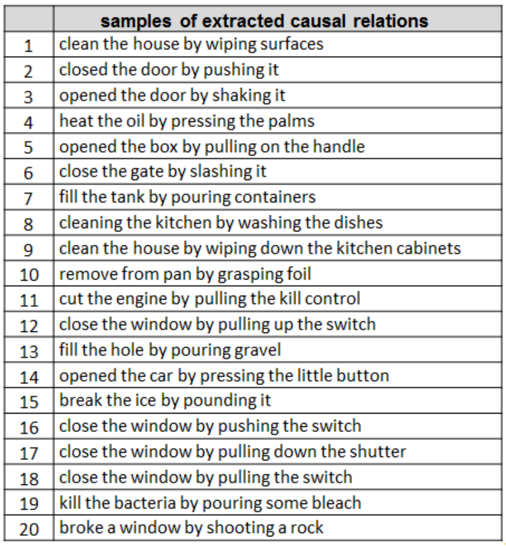

WordNet
Commonsense Knowledge Base
Action Verbs (Va)
Result Verbs (Vr)
Extraction of frequent pairs (Vr+NP)
Extraction of causal relations between action verbs and result verbs
Algorithm of separation
N-gram approach for extraction
Web-search using template (Vr+NP) by (Va+*)

Figure 4: Flowchart of causal relations extraction from text



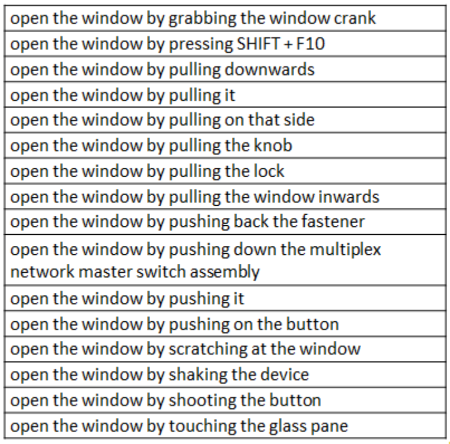
	five the most frequent 3-grams for each of twelve result verbs				
1	breaking new ground	break the ice	broke a window	broke the surface	break the mold
2	clean drinking water	clean the house	clean the air	clean kitchen towel	cleaning the kitchen
3	clear the air	clear the table	clear the area	clear plastic bag	clear the decks
4	closed the door	closed the book	close the window	close the lid	close the gate
5	raised his hand	raised his glass	raise the money	raise the price	raised his arms
6	cut the grass	cut the engine	cut his hair	cut a hole	cut the cake
7	filled the room	fills the screen	fill the space	fill the hole	fill the tank
8	heat the oil	heat olive oil	heat the oven	heat the butter	heat the water
9	killing each other	killed a man	killed his wife	killed the engine	kill the bacteria
10	lifted the lid	lifted his head	lifted a hand	lifted his glass	lift the weight
11	opened the door	open the window	open the gate	opened the box	opened the car
12	remove from heat	remove from oven	removed from office	remove from pan	remove from skillet

Table 1: Most frequent 3-grams for extracted result verbs



	samples of extracted causal relations
1	clean the house by wiping surfaces
2	closed the door by pushing it
3	opened the door by shaking it
4	heat the oil by pressing the palms
5	opened the box by pulling on the handle
6	close the gate by slashing it
7	fill the tank by pouring containers
8	cleaning the kitchen by washing the dishes
9	clean the house by wiping down the kitchen cabinets
10	remove from pan by grasping foil
11	cut the engine by pulling the kill control
12	close the window by pulling up the switch
13	fill the hole by pouring gravel
14	opened the car by pressing the little button
15	break the ice by pounding it
16	close the window by pushing the switch
17	close the window by pulling down the shutter
18	close the window by pulling the switch
19	kill the bacteria by pouring some bleach
20	broke a window by shooting a rock

Table 2: Samples of extracted causal relations



open the window by grabbing the window crank
open the window by pressing SHIFT + F10
open the window by pulling downwards
open the window by pulling it
open the window by pulling on that side
open the window by pulling the knob
open the window by pulling the lock
open the window by pulling the window inwards
open the window by pushing back the fastener
open the window by pushing down the multiplex network master switch assembly
open the window by pushing it
open the window by pushing on the button
open the window by scratching at the window
open the window by shaking the device
open the window by shooting the button
open the window by touching the glass pane

Table 3: Examples of causal relations for the 3-gram "open the window"

7 Conclusion and Further Work

Commonsense inferences allow us to equip and empower cognitive robots with an ability to understand high-level natural language commands (or instructions). We present a method for acquiring the knowledge needed to transform high-level result-verb commands into action-verb commands for further implementation into primitive actions.

In the future, to improve the results and increase the quality of retrieved actions we are planning to:

- improve the instruction for judges to decrease the deviation in evaluation;

- use better NLP tools for POS tagging and parsing;

- develop more elaborated procedure for commonsense inferences, for example, to exclude search results with negation ("don't open the window by throwing the stone") that produce wrong commonsense inferences;

- use metrics for calculation of consistency (reliability) of the results (for example, Krippendorff's alpha coefficient);

- enlarge the set of verbs used for commonsense inferences using resource such as WordNets.

- build multilingual commonsense inferences (starting with Chinese and Indonesian) based on (Bond and Foster, 2013; Bond et al., 2014), (Wang and Bond, 2013).

References

Dilip Arumugam, Siddharth Karamcheti, Nakul Gopalan, Lawson L.S. Wong, and Stefanie Tellex. 2017. Accurately and efficiently interpreting human-robot instructions of varying granularities. In *Proceedings of the Conference on Robotics: Science and Systems.*

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual wordnet. In *51st Annual Meeting of the Association for Computational Linguistics: ACL-2013*, pages 1352–1362.

Francis Bond, Lian Tze Lim, Enya Kong Tang, and Hammam Riza. 2014. The combined wordnet bahasa. In *NUSA: Linguistic studies of languages in and around Indonesia 57*, pages 83–100.

Joyce Y. Chai, Qiaozi Gao, Lanbo She, Shaohua Yang, Sari Saba-Sadiya, and Guangyue Xu. 2018. Language to action: Towards interactive task learning with physical agents. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18).*

Timothy Chklovski and Patrick Patel. 2004. Verbocean: Mining the web for fine-grained semantic verb relations. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Barcelona, Spain.

Herbert H. Clark. 1996. *Using language*. Cambridge University Press, Cambridge, UK.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA, USA.

Christiane Fellbaum and George A. Miller. 1990. Folk psychology or semantic entailment? a reply to rips and conrad. *The Psychological Review*, 97:565–570.

Amy Fire and Song-Chun Zhu. 2016. Learning perceptual causality from video. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 7(2):23.

Maxwell Forbes and Yejin Choi. 2017. Verb physics: Relative physical knowledge of actions and objects. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics: ACL-2017 (Volume 1: Long Papers)*, pages 266–276.

Qiaozi Gao, Malcolm Doering, Shaohua Yang, and Joyce Y. Chai. 2016. Physical causality of action verbs in grounded language understanding. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics: ACL-2016*, pages 1814–1824, Berlin, Germany.

Peter Gardenfors. 2017. *The Geometry of Meaning: Semantics Based on Conceptual Spaces*. MIT Press, Cambridge, MA, USA.

Peter Gardenfors and Massimo Warglien. 2012. Using conceptual spaces to model actions and events. *Journal of Semantics*, 29(4):487–519.

Aliaksandr Huminski and Hao Zhang. 2018a. Action hierarchy extraction and its application. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018). Workshop "Annotation, Recognition & Evaluation of Actions (AREA)"*, Miyazaki, Japan.

Aliaksandr Huminski and Hao Zhang. 2018b. Wordnet troponymy and extraction of "manner-result" relations. In *Proceedings of the 9th Global WordNet Conference (GWC 2018)*, Singapore.

Karin Kipper Schuler. 2005. *VerbNet: A broad-coverage, comprehensive verb lexicon*. Ph.D. thesis. Computer and Information Science Dept. University of Pennsylvania. Philadelphia. PA.

Hadas Kress-Gazit, Georgios E. Fainekos, and George J. Pappas. 2008. Translating structured english to robot controllers. *Advanced Robotics*, 22(12):1343–1359.

Beth Levin. 1992. *English Verb Classes And Alternations*. University of Chicago Press, Chicago, IL.

John P. McCrae, Alexandre Rademaker, Francis Bond, Ewa Rudnicka, and Christiane Fellbaum. 2019. English wordnet 2019 – an open-source wordnet for english. In *Proceedings of the 10th Global WordNet Conference GWC-2019*, Wroclaw, Poland.

Dipendra Kumar Misra, Tao Kejia, Liang Percy, and Saxena Ashutosh. 2015. Environment-driven lexicon induction for high-level instructions. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language*

Processing: ACL-2015 (Volume 1: Long Papers), pages 992–1002, Beijing, China.

Dipendra Kumar Misra, John Langford, and Yoav Artzi. 2017. Mapping instructions and visual observations to actions with reinforcement learning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Malka Rappaport Hovav and Beth Levin. 2010. Reflections on manner/result complementarity. In *Syntax, Lexical Semantics, and Event Structure*, pages 21–38, Oxford University Press, Oxford, UK.

Lanbo She and Joyce Y. Chai. 2016. Incremental acquisition of verb hypothesis space towards physical world interaction. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics: ACL-2016 (Volume 1: Long Papers)*, pages 108–117, Berlin, Germany.

Lanbo She and Joyce Y. Chai. 2017. Interactive learning of grounded verb semantics towards human-robot communication. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics: ACL-2017 (Volume 1: Long Papers)*, pages 1634–1644.

Rion Snow, Daniel Jurafsky, and Andrew Y Ng. 2006. Semantic taxonomy induction from heterogenous evidence. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual meeting of the Association for Computational Linguistics: ACL-2006*, pages 801–808.

Michael Tomasello. 2008. *The Origins of Human Communication*. MIT Press, Cambridge, MA, USA.

Shan Wang and Francis Bond. 2013. Building the chinese open wordnet (cow): Starting from core synsets. In *Proceedings of the 11th Workshop on Asian Language Resources: ALR-2013*, pages 10–18.

Massimo Warglien, Peter Gardenfors, and Matthijs Westera. 2012. Event structure, conceptual spaces and the semantics of verbs. *Theoretical Linguistics*, 38(3-4):159–193.

Terry Winograd. 1972. *Understanding natural language*. Academic Press, Oxford, England.

Jiajun Wu, Erika Lu, Pushmeet Kohli, William T. Freeman, and Josh Tenenbaum. 2017. Learning to see physics via visual deanimation. In *Proceedings of the 31th Annual Conference on Neural Information Processing Systems (NIPS)*.

Rowan Zellers and Yejin Choi. 2017. Zero-shot activity recognition with verb attribute induction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.