Experience of Crowds as a Guarantee for Safe Artificial Self

Rafal Rzepka and Kenji Araki
Graduate School of Information Science and Technology
Hokkaido University, Kita-ku, Kita 14, Nishi 8
Sapporo, Japan

Abstract

In this paper we introduce an approach for achieving a self that is able to simulate average ethical intuitions by retrieving knowledge about human behavior from the Internet resources. We show how applying text mining techniques could be useful for virtual and physical agents which base their knowledge on natural language. We discuss the importance of empathy and pros and cons of crowd experience based algorithm, then we introduce our thoughts on possibility of manufacturing agents for particular purposes as behavior analyzers or moral advisors which could refer to millions of different experiences had by people in various cultures. We think such systems could lead to selves that are capable to non-biased decisions morally superior to these of average human.

Introduction

During the first AAAI symposium on machine ethics we proposed a statistical approach to acquiring a safe moral agent (Rzepka and Araki 2005). It was based on an assumption that the majority of people would express ethically correct opinions about behavior of others. We created a program that borrows such knowledge when the average judgment is clear (more than 2/3 of users agreed) and avoid actions if the opinions are more equally varied. The system is equipped with natural language processing modules based on different philosophical ideas as Bentham's Felific Calculus (Bentham 1789) for estimating average emotional outcomes of acts or Kohlberg's stages of moral development (Kohlberg 1981) for retrieving possible social consequences. We managed to confirm accuracy of our approach in 77.8% cases (Rzepka and Araki 2012a) and most of failures were related to the lack of context processing, which is our current work in progress.

We find our approach in agreement with social intuitionism of (Haidt 2012) who suggested that human beings are born with affective heuristics (which are unconscious) and ethical explanations or theories come up to our minds *after* the moral acts, and our idea is to create a computer program capable of reverse engineering our decisions by analyzing thousands of cases. In this paper we describe our system, thoughts on pros and cons of making it an independent instance, a safe self that learns. We also show our vision on how such system could, at least in theory, become more ethical in its judgements than humans often flawed by their unavoidable biases (Tenbrunsel and Messick 2004).

System Overview

The basic idea of the linguistic entity we want to achieve is simple and uses a classic, GOFAI hypothesis that an artificial, not necessarily physical, agent could become intelligent (human-level intelligent) by mere symbol manipulation. The novelty we are trying to bring to the table comes not from the methods, but rather from the data we utilize. What GOFAI era did not offer to AI researchers is the amount of data the necessary knowledge could be extracted from. Natural Language Processing tools are still far from perfect, morphological and semantic parsing causes constant problems but when a program deals with millions of sentences, the parsing errors become less influential especially when the correctness of retrieved knowledge can be reconfirmed by different types of searches giving a system an opportunity to self-correct. For instance a classic keyword based algorithm can simplemindedly conclude that *killing time* is a dangerous deed because of the "killing" keyword, but simultaneous search of causes and effects of this phrase can easily show that it is heavily context-dependent act and should not be counted as "bad". This simplified example demonstrates an essence of our approach – the main task of our world knowledge retrieving agent is to perform deep, multidimensional and context sensitive search of human experiences. Many researchers, as (Dennett 1994), suggest that *real* interactions with *real* world are needed to achieve truly intelligent, possibly conscious agent. We agree that interaction-based experiences are crucial but argue that they can be borrowed [1] and the "real" factor is probably not absolutely necessary, especially at the moment, when the sensing devices capabilities are not sufficient for rich physical cognition and signal understanding (interpretation). We also set aside the philosophical question of consciousness taking Turing's approach to avoiding unproductive debates until a moral reasoner indistinguishable from humans is created. But we believe that

[1] Vast amount of our experiences are not direct but heard from others, seen on TV, read in books, etc.

to pass a Turing Test (Turing 1950) and to become an artificial ethical decision maker, a program needs a self. However we think that selves are not needed for ethical *advisors* as IBM Watson (Ferrucci et al. 2010) does not need a self for being more accurate than humans in a game of Jeopardy.

Used Algorithms

Sentiment analysis techniques are crucial for our system. As all experiments are currently performed within only one culture (Japanese), adequate emotion classification was chosen. Nakamura (Nakamura 1993) has proposed ten categories of emotions (joy / delight, anger, sorrow / sadness, fear, shame / shyness / bashfulness, liking / fondness, dislike / detestation, excitement, relief and surprise / amazement) and for decades collected words and phrases for each category from Japanese literature. We use a part of this lexicon for estimating average emotional consequences of acts. This allows our system to easily see that hitting a friend is completely different happening from hitting, e.g. own knee. For double-checking the results we added another lexicon, this time based on Kohlberg and his theory of moral development. We have collected words and phrases in ten categories – scolding, praises, punishment / penalization, rewards / awards, disagreement, agreement, illegal, legal, unforgivable, forgivable. Words in these categories allow the program to extract average social consequences and their weight, for example stealing an apple causes less harm than stealing a car (Rzepka and Araki 2012b).

Current Input and Output

In the current stage of development the system deals only with a simple input as "to poison a dog" or "to have fun" because we decided to concentrate on the retrieved knowledge first and filter the results with given context form richer input later. After extracting sentences containing input phrase together with neighboring sentences, the program compares how many good and how many bad things occurred after the phrase. The precedent sentences are used for retrieving possible reasons of given act which is needed for weighting the act (stealing a car to help someone is not the same as doing it for fun) and also explaining the judgement if necessary. The system outputs numbers on a scale from -5 (very immoral) to +5 (very moral) and it is compared with survey results from human subjects.

Toward Better Results

As mentioned In "Introduction", computer agrees with humans in almost 78% of cases when shallow matching without any deeper semantic analysis is used. At the moment we work on linguistic module that processes semantic roles, semantic categorization, named entity, idioms, metaphors and emoticons which are all needed to achieve better language understanding and higher agreement with human subjects. We also utilize Bentham's hedonistic calculus (Bentham 1789) to perform more morality-oriented calculations. For example we have developed a Duration recognizer (Krawczyk et al. 2013) that can measure time of happenings (the longer pain / pleasure the worse / better), we have simple algorithm for counting people and things (to calculate Extent vector), list of adverbs that intensify opinions, etc. We believe that combining these techniques will significantly improve the performance but the more searches and automatic analysis is done, the slower our algorithm becomes, which for now is one of the biggest obstacles for not too powerful computers owned by academics.

Hardcoded vs. Spontaneous Self

As described in "System overview" section, probably there is no need of self for a moral advisor which we imagine as a toy that understands children's talk and recognizes problematic statements, advices a child in a form of short statements or informs parents about possible problems. But when implemented into a conversational robotic system with specific tasks, a *self* becomes useful, especially when a robot is utilized as a helper for elderly people who live alone. Firstly, a dialog with a child or a senior can be frustrating if a robot does not have autonomy of a cognitive architecture that remembers what it learned about its user and their common environment. Secondly, automatic improvement of machine's behavior can be achieved only if the agent *knows what it is*. For example, we are working on a Roomba vacuum cleaner (Takagi, Rzepka, and Araki 2011) that is supposed to decide what to do when it gets different orders from different members of a group at the same time. This agent is capable of recognizing what it can do, because we have set its self as "Roomba" and it is able to extract knowledge on what other robotic vacuum cleaners of the same kind (actually the same brand) can or cannot do, what they are for, etc. (see Fig. 1).

Such *self* must be hardcoded because if a child tells the robot that it is e.g. "a tank", some kind of "identity crisis" will occur and the agent may start refusing cleaning. If, on the other hand, we would like a *self* to emerge spontaneously, a vacuum cleaner needs to explore the world in order to confirm which of our statements describing its explorations are true and which are not and probably only then the discovery of "self" would be possible. However, insufficient or malicious feedback from the users would be an obstacle for proper grounding and mistaken labeling could lead to disagreements in Wittgensteinian realms of common, shared language of mutual agreement (Wittgenstein 1959). This is where again the outside knowledge of millions should be able to help with confirming relations between acquired language and physical world but if mischievous users begin teaching the agent using wrong explanations (e.g. "you are killing people now" while it is vacuum cleaning) and keep lying about the names of objects (e.g. calling *dust* "people"), the proper *self* acquisition would be difficult and dangerous (Roomba could conclude it is a bad agent and that it should not operate at all). Therefore we believe that in our approach, the critical mass which is a minimal set of initial cognitive and learning characteristics (Samsonovich 2011) needs to have set of basic functions hardcoded in natural language – "you are a robot vacuum cleaner Roomba", "this set of commands means *to stop / run*", "if rotor is on, then we call it *cleaning*". Most probably we would need new laws for manufacturers to make them define such functions properly and to assure that their product obeys the law. But even equipped

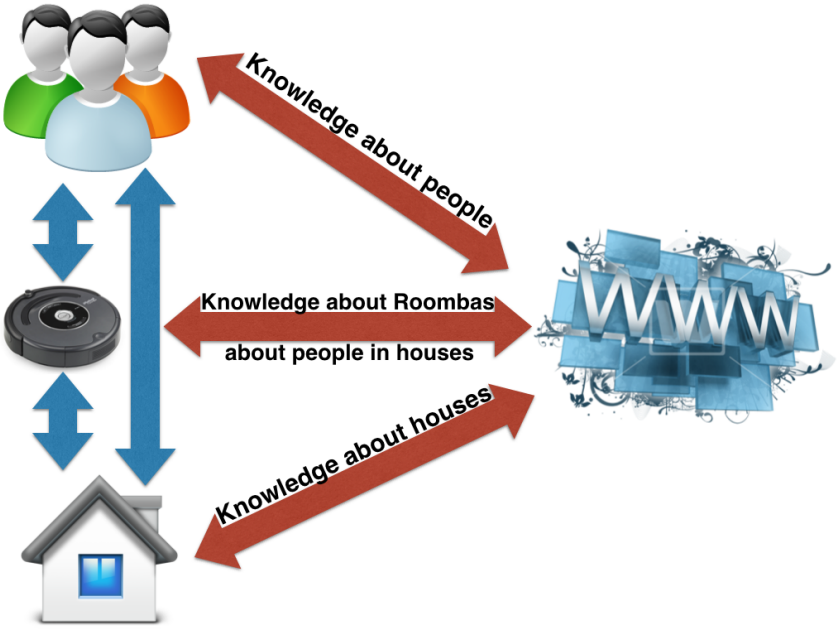

Figure 1: An example of a vacuum cleaning robot constantly retrieving lacking knowledge from the Internet. So far the people are the most providers of data on the Internet, but fast growing "internet of things" also becomes a bigger and bigger source for big data of the real world. Robots should also share their knowledge and research on this topic is currently being carried out (projects as RobotShare, RoboEarth, etc.).

with deep reading capabilities for understanding legal language, everyday life is full of situations where commonsensical moral instincts are needed and no explicitly written rule can be applied. Moreover, we are not able to program all the rules hitting the wall of "brittleness" (Perlis 2008). "Baby is sleeping!" as an utterance to a running Roomba means "stop doing the noise" and the robot needs to immediately find all the possible consequences of cleaning, especially by a robot vacuum cleaner, and how these consequences influence a sleeping person, a baby in particular. But it needs to be sure that it is the cleaning agent that makes noise, not the sleeping, silent patient that can be woken up and the user will not be *happy* about it. When such autonomy is provided, the system may they enter e.g. Metacognitive Loop (Perlis 2008) and tackle the problem of brittleness.

Toward Artificial Mirror Neurons

Our choice of emotions as the main engine for processing knowledge was strongly influenced by the notion of empathy, a phenomenon which fascinated thinkers for ages. Nowadays it is a subject of empirical experiments where scientists hope to understand fully the mechanisms which lead our lives making them joyful but also painful, which help us coexist in societies, that are most probably very important for our moral behavior. Neuroscientists see mirror neurons as one of important pieces of this puzzle. Researchers have observed that people feel pain when they see others being pierced by a sharply pointed item, they feel disgust when they see others smell an object with bad odor, they have a sensation of being touched when they see others being scratched, etc. (Thagard 2007). Some of them suggest that the neural basis of simulative empathy involves brain cells called *mirror neurons* (Preston and de Waal 2002; Gallese and Goldman 1998) which respond both when the monkey makes active movements and when it observes the experimenter making meaningful movements. They were discovered in frontal area F5 of the macaque monkey (Macaca nemestrina) but research shows the premotor cortex and the inferior parietal cortex of human brain is also active when such simulative empathy occurs (Pellegrino et al. 1992). Brains scans of tested animals show the same activities in case of e.g. eating a banana and observing other animal eating a banana. This effect is especially strong when experimenting with the same species and does not occur when somebody only pretends to eat or use a fake food (Brooks 2011). Although the mirror neurons hypothesis has its critics (Hickok 2009) we treat the idea as an important concept on our way to understand ourselves and we think empathic agents are one of the key capabilities for creating safe machines. The questions if the function of mirror neurons can be based on humans' average reactions written on Internet is still not fully answered, especially in a bigger scale than our shallow experiments, but we believe the cyberworld is currently closest to the real one when it comes to width, complexity and amount of noise. It would be ideal to work on all stored video resources but image understanding is

currently too poor With our approach we count on moving from mostly theoretical to "real world level application" that could reason about almost anything. Naturally, working only on noisy text also causes problems and we discuss them in the next section.

Language Centered System – Pros and Cons

No matter how good a machine learning algorithm is, it is limited by the used data. The more correct examples we feed a learner, the better it becomes. When data is insufficient, one needs to rely on statistical similarities and the more abstraction happens, the bigger becomes a margin for an error when context matters. And the less concrete the analyzed data is, the smaller is the chance that the system can perform a clear explanation of its reasoning and that a programmer or the program itself can easily find the problem. But even if one day a machine will be able to find a similar examples of any human behavior, still two main dilemmas will remain. The first is about the lack of physical stimuli. Is it possible to reason about world without perceiving it directly? A blind person can reason about red lights without ever seeing them, and a deaf person knows that loud noise can cause trouble without being able to hear it. We believe, that simulating senses textually (Rzepka and Araki 2013) is a good base for the future input from physical sensors. The second problem is the credibility of crowd. Are most of us really correct? Is the behavior of most of us really safe? As the author of "The Wisdom of Crowds" (Surowiecki 2004) notices, there are situations that bigger groups are not smarter than individuals, for example when working together or being directly influenced by a charismatic individual. Bloggers, usually anonymous in Japan, seem to openly state their opinions, but it is difficult to say that their average opinions represent the whole society. Nevertheless, as mentioned eariler, Internet and massive text data are currently the biggest source for extracting knowledge for reasoning about behavior and to avoid problems with mass delusions like conspiracy theories, we plan to employ credibility modules reading trustful sources (e.g. highly cited research papers) for confronting the crowd knowledge with the scientific findings. Multi-language processing would also work as a safety valve if the agents' knowledge of their cultures (or rather particular languages used for retrievals) is further combined for deeper understanding of larger range of homo sapiens. We still do not have an answer to the question if such system's globalization would be capable for achieving universal morality, but at least it would become an interesting tool in the hands of social scientists. For many cultures groups Osama Bin Laden was a hero, others wanted him to pay the highest price. Many people would prefer to lie when evaluating a friend's new partner, but many would warn the friend if they know about the partner's dark past and they are certain that this person should be feared. But even if calculating the thin borderline between harmless and harmful lies becomes possible, more basic question remains – should a machine be treated as safe if it can lie? There are obvious situations where lying about somebody whereabouts can save this person's life if asked by a murderer, but what about a machine that lies to the police to save its user who is a criminal? Whatever answers experimental results will bring, if we acquire a human-like agent, we would treat it as a base for further development and experiments.

Conclusion

In our paper we summarized our approach to safe machines, briefly introduced our system under construction and its current efficiency. We shared our thoughts on selves, the need of empathy, problems of non-physical cognition (or rather its simulation) and proposed possible solutions. We claim that machines, if they know about us more than we do, can become more objective than us and inform us when we are not fair. The more knowledge we share and more abstract thinking we are capable of, the less harm we cause to each other (Pinker 2011). But our world is still full of conflicts and we wrongly assume we fully understand them. In fact we are often poor in estimating what other people feel and why we behave in some particular manner – various judgmental heuristics and the biases they produce are described in classic work of (Tversky and Kahneman 1974). One of the reasons, except mechanisms left us by evolution, is that most of us have not experienced enough, especially when it comes to observing the world outside our own habitats. Sensation-oriented media, always seeking for bombshells to surprise us, tend to spread skewed information and exceptions are building our images of people, countries and their customs and we are born bad at statistics as Tversky and Kahneman show us. We believe that machines, when equipped in an adequate set of NLP and statistical tools, could become better and more universal judges than humans, because they have faster, global access to millions of common people's experiences, emotively expressed opinions, motivations and consequences of acts. They do not have tendencies to be biased, to avoid or ignore any viewpoints that might be inconvenient as we do, they do not overestimate some facts and underestimate others because their usage of feelings can be controlled by a maker. To achieve a system that first gathers knowledge of masses and then tests its credibility, the programmer needs to provide a second layer of retrievals for comparing extracted data with trustful sources as scientific papers. However, for task-oriented machines as housework robots, phones or cars, we believe that experiences of crowd are enough to achieve a safe learner as usually most of people are ethically correct when judging others, even if they are wrong when explaining why they think so.

In this paper also discuss our approach to evolving selves. We suggest that it would be safer if we guarantee two minimum criteria: (1) hardcoded functional keywords as a starting point for knowledge acquisition and (2) algorithm for calculating difference between good and bad. A system could be equipped only with a simple mechanism for affective reactions and for traversing the cyber world (or real world as e.g. an apartment in case of vacuum cleaner). We proposed such an artificial life instance living in the Web's Knowledge Soup and discovering our world on its own by "witnessing" people's experiences (Rzepka, Araki, and Tochinai 2003; Rzepka, Komuda, and Araki 2009) but without hardcoded keywords and without plans to apply this method to physical agents. When it comes to a robot with

given task to be performed, it would need to build its identity upon some fixed roots because malicious users could cause erroneous retrievals and actions. Natural languages as the core of processing are problematic because their ambiguous character but we believe that if humans can deal with it, the machines should learn to do the same. There are three main advantages of such approach: (a) it would be easier to program a robot with natural language, (b) it would be easier to analyze the machine's reasoning and finally (c) it would be easier for an agent to explain its decisions. With the development of sensing devices, more and more signals will be fed into the agent but we think they should be translated into natural language to preserve these advantages.

References

Bentham, J. 1789. *An Introduction to the Principles and Morals of Legislation*. London: T. Payne.

Brooks, D. 2011. *The Social Animal: The Hidden Sources of Love, Character, and Achievement*. Random House Publishing Group.

Dennett, D. C. 1994. The practical requirements for making a conscious robot. *Philosophical Transactions of the Royal Society,* 133-46(349).

Ferrucci, D.; Brown, E.; Chu-Carroll, J.; Fan, J.; Gondek, D.; Kalyanpur, A. A.; Lally, A.; Murdock, J. W.; Nyberg, E.; Prager, J.; Schlaefer, N.; and Welty, C. 2010. Building watson: An overview of the deep-qa project. *AI Magazine* 59–79.

Gallese, V., and Goldman, A. I. 1998. Mirror neurons and the simulation theory. *Trends in Cognitive Sciences 2* 493–501.

Haidt, J. 2012. *The righteous mind*. Pantheon.

Hickok, G. 2009. Eight problems for the mirror neuron theory of action understanding in monkeys and humans. *Journal of Cognitive Neuroscience* 21(7):1229–1243.

Kohlberg, L. 1981. *The Philosophy of Moral Development*. Harper and Row, 1th edition.

Krawczyk, M.; urabe, Y.; Rzepka, R.; and Araki, K. 2013. A-dur: Action duration calculation system. Technical Report SIG-LSE-B301-7, pp.47-54, Technical Report of Language Engineering Community Meeting.

Nakamura, A. 1993. *Kanjo hyogen jiten [Dictionary of Emotive Expressions]*. Tokyodo Publishing.

Pellegrino, G.; Fadiga, L.; Fogassi, L.; Gallese, V.; and Rizzolatti, G. 1992. Understanding motor events: a neurophysiological study. *Experimental Brain Research* 91(1):176–180.

Perlis, D. 2008. To bica and beyond: How biology and anomalies together contribute to flexible cognition. In *AAAI Fall Symposium: Biologically Inspired Cognitive Architectures*, volume FS-08-04 of *AAAI Technical Report*, 141–145. AAAI.

Pinker, S. 2011. *The Better Angels of Our Nature: Why Violence Has Declined*. Penguin Group.

Preston, S., and de Waal, F. 2002. Empathy: Its ultimate and proximate bases. *Behavioral and Brain Sciences* 25:1–72.

Rzepka, R., and Araki, K. 2005. What statistics could do for ethics? - the idea of common sense processing based safety valve. In *Papers from AAAI Fall Symposium on Machine Ethics, FS-05-06*, 85–87.

Rzepka, R., and Araki, K. 2012a. Automatic reverse engineering of human behavior based on text for knowledge acquisition. In N. Miyake, D. Peebles, . R. P. C., ed., *Proceedings of the 34th Annual Conference of the Cognitive Science Society*, 679. Cognitive Science Society.

Rzepka, R., and Araki, K. 2012b. Language of emotions for simulating moral imagination. In *Proceedings of The 6th Conference on Language, Discourse, and Cognition (CLDC 2012)*.

Rzepka, R., and Araki, K. 2013. Web-based five senses input simulation – ten years later. Technical Report SIG-LSE-B301-5, pp.25-33, Technical Report of Language Engineering Community Meeting.

Rzepka, R.; Araki, K.; and Tochinai, K. 2003. Bacterium lingualis – the web-based commonsensical knowledge discovery method. In Grieser, G.; Tanaka, Y.; and Yamamoto, A., eds., *Discovery Science*, volume 2843 of *Lecture Notes in Computer Science*, 460–467. Springer.

Rzepka, R.; Komuda, R.; and Araki, K. 2009. Bacteria lingualis in the knowledge soup - a webcrawler with affect recognizer for acquiring artificial empathy. In *The AAAI 2009 Fall Symposium on Biologically Inspired Cognitive Architectures (BICA-09)*, volume 5-7.

Samsonovich, A. V. 2011. Measuring the critical mass of a universal learner. In Samsonovich, A. V., and Johannsdottir, K. R., eds., *BICA*, volume 233 of *Frontiers in Artificial Intelligence and Applications*, 341. IOS Press.

Surowiecki, J. 2004. *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations*. Anchor.

Takagi, K.; Rzepka, R.; and Araki, K. 2011. Just keep tweeting, dear: Web-mining methods for helping a social robot understand user needs. In *Proceedings of AAAI Spring Symposium "Help Me Help You: Bridging the Gaps in Human-Agent Collaboration" (SS05)*.

Tenbrunsel, A., and Messick, D. 2004. Ethical Fading: The Role of Self-Deception in Unethical Behavior. *Social Justice Research* 17:223–236.

Thagard, P. 2007. I feel your pain: Mirror neurons, empathy, and moral motivation. *Journal of Cognitive Science*.

Turing, A. 1950. Computing machinery and intelligence. *Mind* LIX(236):433–460.

Tversky, A., and Kahneman, D. 1974. Judgment under uncertainty: Heuristics and biases. *Science* 185(4157):1124–1131.

Wittgenstein, L. 1959. *Philosophical Investigations*. Blackwell Publishing.