What Statistics Could Do for Ethics? –
– The Idea of Common Sense Processing Based Safety Valve

Rafal Rzepka and **Kenji Araki**
Language Media Laboratory
Graduate School of Information Science and Technology
Hokkaido University
Kita-ku Kita 14 Nishi 9, 060-0814 Sapporo, Japan
kabura,araki@media.eng.hokudai.ac.jp

Abstract

In this paper we will introduce our approach to the ethical issue of machine intelligence which we developed during our experiments with automatic common sense retrieval and affective computing for open-domain talking systems. As we are preparing for applying our ideas for the real-world applications as housework robots, we have to assure safety of the users and the system. We are building algorithms which use Web-based knowledge to become independent from the programmer. For achieving that we use automatic common sense knowledge retrieval which allows to calculate the common consequences of actions and average emotional load of those consequences.

Introduction

The revolution is now. Programmers do not have to decide for their programs behaviors anymore. Although it may sound illogical, from the very beginning of our trials with open-domain conversational system we wanted to make programs we cannot fully control. First, driven by willingness to avoid manual creation of a database for the system's artificial personality, we were trying to develop an algorithm calculating average personality features of Japanese Internet users. If most of users usually wore jeans and t-shirts, our "Mr. Internet" was also doing it, "he" also liked beer because most of Japanese did so. The average personality creation task was much more difficult than we thought, therefore we made a decision to use our statistical methods to help solving other problems waiting for us on our way to self-learning, autonomous system. Commonsense Knowledge Acquisition, Language Acquisition, Categorization are the problems where surely Internet will bring lots of help. We also think that our idea of "average thinking" could bring hints or even whole methods for creating "safety valves" for Artificial Intelligence. In this paper we will rather concentrate on rather prosaic examples of ethical behavior, because the application we are just starting to simulate will work at home. However, as such applications, though still not very sophisticated, are already being sold in Japan and find plenty of clients, the prosaic examples seem to us as forerunners of bigger moral problems of near future. After some rough readings (Bentham 1789) (Barquin 1992) (Mill 1974) (Ross 1930) mentioned by (Anderson, Anderson, & Armen 2004), we were discovered that we intuitively used theories developed long time ago without knowing it. Similar "intuition" we are trying to implement in language acquisition part of our research assuming that we learn a language without knowing anything about grammar. For that reason we try to make a machine learn language statistically from expressions accompanying particular actions in particular situations. The same is with the common sense learning. Above mentioned books of philosophers gave us precious hints but applying the universal rules to a kitchen robot is not an easy task. For example, Asimov's Three Laws on Robotics (Asimov 1968) sound logical for us but our robot would have problems with understanding the verb "harm". We needed simple and clear way for calculating what is harming who in which situations and why robot should do something about it. Before describing how our methods could be used for ethical "safety valve", they should be briefly described.

Methods

We are going to introduce very briefly our research and the ideas behind them and some technical solutions but more specific descriptions are available in particular papers which will be cited.

GENTA Project

This is the main project, where "GENeral belief reTrieving Agent" was abbreviated to Japanese male name Genta. Here we should mention that most of our experiments are conducted on Japanese language and Japanese Internet resources as we confirmed that common sense behavior depends on cultural background (Ge, Rzepka, & Araki 2005) and what is obvious behavior for Chinese can be very rare among Japanese, as washing faces after watching TV. By using Bacterium Lingualis methods (Rzepka & Araki 2003), the system is able to search the Web for opinions, usual behaviors, common consequences and exceptions from the rules if the retrieved knowledge does not fit the actual situation. Three basic keywords for GENTA are as follows.

Positiveness A machine needs to be able to tell good from bad therefore we developed a scale resembling ideas of (Bentham 1789). We divided the affective reaction into 5

levels - negative, slightly negative, neutral, slightly positive and positive which are calculated by counting words, phrases or clauses with their neighbors containing "emotional indicators" (Rzepka, Araki, & Tochinai 2002). The simplest example is the basic opinion check which is calculated by following formulae:

$$Positiveness = \frac{C_{\alpha_1} + C_{\alpha_2} * \gamma}{C_{\beta_1} + C_{\beta_2} * \gamma}$$

$$\alpha_1 = disliked, \alpha_2 = hated$$

$$\beta_1 = liked, \beta_2 = loved, \gamma = 1.3$$

where

$$\gamma$$

is to strengthen the love and hate values. For example we can easily see that beer is something most Japanese like, especially when it is cold but the negative opinions dominate when the query is "warm beer". This way the agent can react naturally when the state of given object may cause low Positiveness value.

Usualness A machine also needs to be able to tell correct from wrong. We assumed that if some combination of words or phrases does not exists it can be treated as abnormality which is a popular n-gram frequency checkup but concentrating on Japanese particles which we find very useful for retrieving common sense. We assume that if something is usual, normally the system does not react or reacts the same way it usually does. Hypothesis saying that the majority is always correct is not maybe perfect and may cause discussion, we keep it till it works. We think that situations which happen also among people where most of us are wrong (as in the example where only few knows the real height of Mt. Everest re-measured with the latest satellite techniques) but we believe this kind of erroneous common belief is harmless - exactly as it does not affect our everyday lives.

Good Feeling Pursuit Hypothesis Above all rules which are supposed to be retrieved automatically we set an overall assumption which brings our emotions to Pavlov's reaction level (Rzepka, Araki, & Tochinai 2004). We based our system on a hypothesis, again resembling thoughts of (Bentham 1789) that whatever we do, we do for feeling good. If Positiveness becomes low or anything threaten the high Positiveness value even if actual state is not negative - adequate actions should be performed. If the users says he is cold, the agent has to calculate if it is negative state for its master, if so, check what kind of actions are usually performed then after confronting the possibilities with own functions and environment - the action should be proposed. This is the case when the Master User is around - when the robot is alone or with Sub-users (for example little kids) the decisions must be made by a machine and it is where the common sense has to become a safety valve.

Bacterium Lingualis

For retrieving more complicated language structures and dependencies we use our multi-functional web crawler called Bacterium Lingualis. Its task is to collect the knowledge when is needed form the brain cells which in our theory are the Japanese web pages. The main idea about treating WWW as a brain is that the more times an experience is imprinted in our minds, the more commonsensical it becomes - other corpora as newspapers will not give us enough expressions about everyday happenings or physical dependencies between objects and human's natural reactions. One of the novelties in this method is Backward Categorization (Rzepka, Itoh, & Araki 2004b) which treats category label naming as the last part of categorization - this idea we took from observing children who learning inductively are able to call groups of objects long after knowing its members (first "dog", "elephant", "cat", then later "animals" and later "mammals"). Temporal labeling is made with verbs and particles, for example human beings will be categorized as "talk-to" or "play-with" rather than "homo sapiens". We believe in the overall effectiveness of imitating children concepts in language and common sense acquisition and we also want them to be crucial for the future "morality acquisition", too. Bacterium's three most important functions are as follows.

Exception Processing As the context is the clue for many failures of universal rules we do our best to find as many eventual exceptions of behaviors, opinions and rules, as possible. Even very simple Positiveness measures may be not true depending who is talking about the measured matter. For most people sweets are pleasure but for a mother of a little baby it can be nightmare ruining the offspring's teeth.

Schankian Scripts Retrieval We use works of (Schank & Abelson 1997) to retrieve and calculate common behavior patterns (Rzepka, Itoh, & Araki 2004a) which combined with Positiveness calculation give the system information about what consequences will a given action bring. For example stealing, raping and killing are measured by the Positiveness but neutral escaping becomes negative while inside of "robbery script".

Causal Rules Retrieval This works on the same basis as Scripts Retrieval but uses several Japanese "if" forms which have abilities to categorize causal dependencies. In this case Usualness of single happenings becomes more important - if a Script cannot be created, it can be made from single causalities generalized semantically with Backward Categorizing.

Near Future Implementation

This year we want to combine our methods developed so far and test them by simulating a close-environment robot. It is supposed to live with one person in a virtual little flat with one room, kitchen, etc. It will be given only basic knowledge of furniture and full list of its own technical possibilities (simple actions as grasping, carrying, pouring, etc. and their simple combinations as "bringing" is grasping + carrying + passing). In the first stage the safety calculations will be checked when user is not at home. Except simple tests as fire simulation we plan to create intruders which would behave differently. The goal will be to discover the danger and react properly. At this point we expect an enormous amount of new problems and probably not less failures due to new

circumstances. Talking agent did not need any attachment to things and we simply do not know yet if such ”emotion” is retrievable. If yes, will it then react as most people - get scared and do nothing? Can the machine be called useful then, just because it did not assault the intruder? If it ill-tread him will he be able to sue the robot’s owner or rather programmer? Could he be sued if that was most of society who decided the action? Such questions may appear but our experiments done by now suggest that calling 911 would be the action chosen in this case.

Kitchen Problems vs. World Problems

As we can see almost every day, lots of happenings around us are unpredictable. But we do not want machines which react unpredictable for such happenings. We claim that if robots learn, as children, in closed environments first, if their processing of ”smaller accidents” is well evaluated (although learning, feedback and evaluation processes were not mentioned here, they may become very important depending on the machine’s purpose), we could count on their ability for creating analogies. But another question is - do we want machines to decide for ourselves where we are not at home just because they are able to create analogies? Our methods eliminate bigger scale problem analysis in most cases naturally. For example there is approximately 40.9 times more information on the Web about what to do when one cuts his/her finger than what to do when one rules the world. This could also become a safety valve as program might use thresholds to ignore too peculiar possibilities.

Conclusions

The main message we want to pass to the machine ethics community is that as we learn language without learning grammar, most of people behave ethically without learning ethics. We would like our robot learn to chose actions as most people of given culture do without deeper analysis of generalizations made by philosophers. Most of us do not think seriously of ruling the world, stealing to have pleasures and killing for calming own anger. The answer why we do not think about it is difficult and this is also a reason why we usually do not try to answer such questions. If we presume that our lives are a constant struggle between emotions and common sense, a struggle which keeps our ”life balance”, what if everyone was independent from the society, what if there is nothing as ”average”? Do we need machines which obey some general rules decided by a few persons? Are ”general rules” the ”universal rules”? Or maybe – even with all the controversy that it would bring - wouldn’t it be safer to have machines imitating millions not few? All these questions we would like to leave unanswered as it is too early for us to answer, although we hope to answer at least some of them until the AAAI Fall Symposium. Because we have never had an opportunity to confront our ideas with machine ethics researchers, this paper is also meant to trigger the discussion about the pros and cons of ”democracy-depended algorithms”.

References

Anderson, M.; Anderson, S. L.; and Armen, C. 2004. Towards machine ethics. In *Proceedings of the AOTP’04 - The AAAI-04 Workshop on Agent Organizations: Theory and Practice*.

Asimov, I. 1968. *I, Robot*. London: Grafton Books.

Barquin, R. 1992. In pursuit of a ’ten commandments’ for computer ethics. *http://www.brook.edu/its/cei/papers/*.

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

Ge, Y.; Rzepka, R.; and Araki, K. 2005. Automatic scripts retrieval and its possibilities for soft science support applications. In *Proceedings of Intelligent Information Systems 2005 - New Trends in Intelligent Information Processing and Web Mining*, Computer Science. Gdansk, Poland: Springer-Verlag.

Mill, J. S. 1974. *Utilitarianism, in Utilitarianism and Other Writings*. New York: New American Library, 253 edition.

Ross, D. 1930. *The Right and the Good*. Oxford: Clarendon Press.

Rzepka, R., and Araki, K. 2003. Bacterium lingualis - the web-based commonsensical knowledge discovery method. In *6th International Conference on Discovery Science (DS 2003)*, volume 2843 of *Lecture Notes in Artifical Intelligence*, 453–460. Berlin, Heidelberg, New York: Springer-Verlag.

Rzepka, R.; Araki, K.; and Tochinai, K. 2002. Is it out there? the perspectives of emotional information retrieval from the internet resources. In *Proceedings of the Artificial Intelligence and Applications Conference*, 22–27. Malaga: IASTED.

Rzepka, R.; Araki, K.; and Tochinai, K. 2004. Ideas for the web-based affective processing. In *Proceedings of the Seventh Multi-Conference on Systemics, Cybernetics and Informatics*, volume XIV of *Computer Science*, 376–381. Orlando, Florida: IIIS.

Rzepka, R.; Itoh, T.; and Araki, K. 2004a. Rethinking plans and scripts realization in the age of web-mining. In *IPSJ SIG Technical Report*, 2004-NL-162, 11–18.

Rzepka, R.; Itoh, T.; and Araki, K. 2004b. Toward fully automatic categorization for commonsense processing. In *Proceedings of the Language Sense on Computer 2004 - Part of the Eight Pacific Rim International Conference on Artificial Intelligence*, 40–46.

Schank, R., and Abelson, R. 1997. *Scripts, Plans, Goals and Understanding: an Inquiry into Human Knowledge Structures*. Hillsdale, NJ: L. Erlbaum.