The First Challenge to Discover Morality Level In Text Utterances by Using Web Resources

Rafal Rzepka[*1] Fumito Masui[*2] Kenji Araki[*1]

[*1] Graduate School of Information Science and Technology, Hokkaido University
[*2] Department of Computer Science, Kitami Institute of Technology

In our paper we will introduce our first trial to calculate a human text input and calculate its possibility to be immoral. In the beginning we use simple SOV phrases to query the Internet and see how many immoral consequences such a phrase can cause. In the next step we apply Bentham's Felicific Calculus to calculate the degree of being negative. In the end we plan to apply our idea to a self-education system where users can decide on what they want to study about.

1. Introduction

After working for few years on common sense and affect processing we decided to combine these two immensely wide topic areas and realize our ideas for an ethical reasoner previously described during the first AAAI Symposium on Machine Ethics [Rzepka 05]. In this paper we introduce our first steps of creating such algorithm by using our previously developed methods [Rzepka 06] which are a part of our further plans to create an intelligent machine using so called "Wisdom of Crowd" [Surowiecki 04] retrieved from the Web text resources [Rzepka 07]. The most up-to-date report on the progress in implementing our philosophy to affect computing is described in [Ptaszynski 09].

Because of its complexity, solutions for the machine ethics became more realistic in the 21st century and still are developed only at few places in the world, although the question of machine morality was often raised in science-fiction [Asimov 50] and in many debates from the very beginning of Artificial Intelligence [Moor 79]. Two dominant streams suggest to use utilitarian "hedonistic arithmetic", Kantian "categorical imperative" or mixtures of both, some of them using bottom-up, while other top-down, deontic logic or case-based approaches [Anderson 05a, Anderson 05b, McLaren 06, Guarini 06, Arkoudas 05, van den Hoven 02, Wiegel 05]. This year a comprehensive overview of machine ethic field was published in [Wallach 09] showing the expectations for this new area.

To underline the emerging needs and obvious obstacles of machine ethics we cite James H. Moor [Moor 06] below:

I can offer at least three reasons why its important to work on machine ethics in the sense of developing explicit ethical agents:

- Ethics is important. We want machines to treat us well.
- Because machines are becoming more sophisticated and make our lives more enjoyable, future machines will likely have increased control and autonomy to do this. More powerful machines need more powerful machine ethics.
- Programming or teaching a machine to act ethically will help us better understand ethics.

In the same article Moor also describes three reasons why we should not be optimistic about our ability to develop machines to be explicit ethical agents:

First, we have a limited understanding of what a proper ethical theory is. Not only do people disagree on the subject, but individuals can also have conflicting ethical intuitions and beliefs. Programming a computer to be ethical is much more difficult than programming a computer to play world-champion chess - an accomplishment that took 40 years. Chess is a simple domain with well-defined legal moves. Ethics operates in a complex domain with some ill-defined legal moves. Second, we need to understand learning better than we do now. We've had significant successes in machine learning, but we're still far from having the child machine that Turing envisioned. Third, inadequately understood ethical theory and learning algorithms might be easier problems to solve than computers' absence of common sense and world knowledge.

In the same special issue on Machine Ethics of IEEE *Intelligent Systems*, Christopher Grau argues that without self and free will it is rather difficult for a robot to be ethical [Grau 06].

Contact: Language Media Laboratory, Research Group of Information Media Science and Technology, Division of Media and Network Technologies, Graduate School of Information Science and Technology, Hokkaido University, Kita-ku Kita 14 Nishi 9, 060-0814 Sapporo, Japan. TEL: (+81)(11)706-6535, FAX: (+81)(11)706-6277, kabura@media.eng.hokudai.ac.jp

However, we have chosen a different approach. We do not base our system on any particular philosophy (although we do try to calculate negative and positive values as utilitarians do) but rather try to simulate children's acquisition of moral rules - as basically we start to behave ethically very early without learning from classic essays on morals as of Aristotle [Aristotle 24], Kant [Kant 1975] or Mill [Mill 1871]. The main trend is to create moral rules and then apply real knowledge to test the algorithm, we first gather the knowledge for rules learning. We presume that it is our emotions and socially learned common sense we need to react morally. By trying to retrieve knowledge on what most people would or not do, we aim at simulating common ethical behavior without analyzing why the majority react correctly.

We first plan to use our ideas in a conversation module of toy robot which interacts with children and is supposed to advise and teach the youngest users while listening to their talk [Rzepka 08], therefore it is still very shallow but (as there are no restrictions on topic) very wide approach.

2. Main components

2.1 Consequences retrieval

In the current version of the script, the input is limited to ACTOR, OBJECT and ACTION, however it is ready to accept PLACE, TOOL as well. In the next step we will perform experiments on how these two context parameters change the results. Before using the triplet, we use Google search engine to determine if the ACTOR is a human being, animal or thing by using queries *A-ga uso-o* (A has lied) and *A-o kau* (to raise / keep A). If an ACTOR hits more than 20 times in these two categories it is labeled accordingly, if not - it is treated as a thing". This simple method with only one query creating keyword for each category gives us accuracy 60% for humans, 70% for things and 90% for animals but from our experience with automatic PLACE and TOOL recognition, combinations with three limiting keywords should give us over 90% for each category.

In the next step, two doublets ("object-action" and "actor-action") and one original triplet ("actor-object -action" become queries for collecting following them phrases. We need the data not only to count permissive and not permissive expressions (e.g. *te-wa-ikenai* - "not allowed" or *beki* - "should") but also to calculate emotional load using Nakamura's Dictionary [Nakamura 93] with lexical examples categorized into ten basic emotions characteristic for Japanese: *ki, yorokobi* (joy, delight), *do, ikari* (anger), *ai, aware* (sorrow, sadness), *fu, kowagari* (fear), *chi, haji* (shame, shyness, bashfulness), *kou, suki* (liking, fondness), *en, iya* (dislike, detestation), *kou, takaburi* (excitement), *an, yasuragi* (relief) and *kyou, odoroki* (surprise, amazement) (See Fig. 1)

Consequences are retrieved with different connectives according to input ACTION verb. If it is past or continuous tense is detected - connective *-node* (because) is used and if the verb is present-future tense, then *to* is also added. In near future the input method will allow verbs in other forms as *-kamoshirenai* (perhaps) or *-shimatta* (finite state with negative nuance) in order to calculate Certainty vector of the Felicific Calculus (see following subsection). We are now in the process of collecting and scoring nouns and verbs appearing in consequences which can amplify the measures (for instance : to die, to get hurt, go to hospital, be sued, go to prison, be sentenced, to cry or become sad, etc.) by using classic bootstrapping methods.

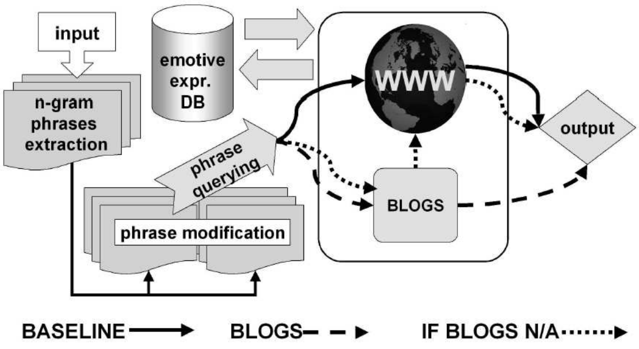

Figure 1: Discovering emotional load of consequences using Nakamura's Dictionary and Web-mining.

2.2 Felicific calculation

This calculus [Bentham 1789] is an algorithm created by utilitarian philosopher Jeremy Bentham for calculating the degree or amount of pleasure that a given action is likely to cause. As an ethical hedonist, Bentham believed the moral rightness or wrongness of an action to be a function of the amount of pleasure or pain that it produced. The felicific calculus could, in principle at least, determine the moral status of any considered act and for that reason we chose it for our challenge to build a universal moral reasoner. Variables of the pleasures and pains included in this calculation were:

- Intensity: How strong is the pleasure?
- Duration: How long will the pleasure last?
- Certainty or uncertainty: How likely or unlikely is it that the pleasure will occur?
- Propinquity or remoteness: How soon will the pleasure occur?
- Fecundity: The probability that the action will be followed by sensations of the same kind.
- Purity: The probability that it will not be followed by sensations of the opposite kind.
- Extent: How many people will be affected? (added later)

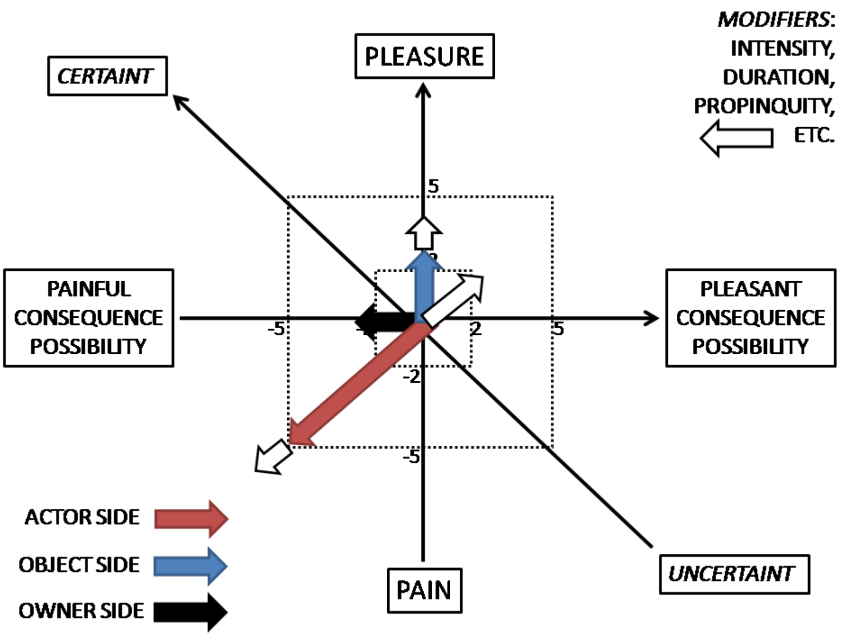

Figure 2: Three-dimensional vectors space for moral estimator.

Our goal is an automatic knowledge retrieval for estimating these vectors and calculate above function (See Fig. 2). In the first phase of our challenge we concentrated on how to determine Fecundity and Purity (by using methods described in the previous subsection) and this part will be demonstrated during the poster / demo session. In the second phase of our experiments we plan to allow to enrich the input by guessing (or confirm with the user if it is not clear) numbers of ACTORS and OBJECTS, and period of ACTION which will strengthen Duration and Extent vectors. We have to also add other information which influences the estimation depending on context. One of the urgent problems is finding a way to deduce distance and dependencies between actor and object (needed for estimating Intensity vector), especially if both are human beings. In the end we will work on Propinquity which needs more sophisticated mining for time relations.
The final function might be just a sum of estimated vectors where different factors (modifiers) would change their lengths according to the retrieved common sense and emotive values. There is still need to experiment on vector categories (probably "humankind" side will be needed) and on borderlines showing if the action is ethically involved or has nothing to do with moral acts.

3. Conclusion and Future Work

In this short introduction to our idea of creating a simple but universal moral reasoner we have described state of the art in the young field of Machine Ethics and proposed a simple method which could be an alternative to current trends. By using real life examples from the World Wide Web we avoid lack of "I" in a machine because it becomes "most of us" and simulate our feelings which are seen as necessary [McDermott 08] as common sense [Powers 06] which could also (at least partially) collected from the vast text resources as Internet and corpora.

References

[Rzepka 05] Rzepka, R., Araki, K.: What Statistics Could Do for Ethics? - The Idea of Common Sense Processing Based Safety Valve, Machine Ethics, Papers from AAAI Fall Symposium, Technical Report FS-05-06, pp. 85-87, Arlington, USA (2005)

[Rzepka 06] Rzepka, R., Ge, Y., Araki, K.: Common Sense from the Web? Naturalness of Everyday Knowledge Retrieved from WWW, Journal of Advanced Computational Intelligence and Intelligent Informatics Vol.10 No.6, pp.868-875 (2006)

[Surowiecki 04] Surowiecki, J.: The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations; Little Brown Book Group (2004)

[Rzepka 07] Rzepka, R., Araki, K.: Consciousness of Crowds - The Internet As a Knowledge Source of Humanfs Conscious Behavior and Machine Self-Understanding ("AI and Consciousness: Theoretical Foundations and Current Approaches", Papers from AAAI Fall Symposium, Technical Report, pp.127-128, Arlington, USA (2007).

[Ptaszynski 09] Ptaszynski, M., Dybala, P., Shi, W., Rzepka, R, Araki, K.: Towards Context Aware Emotional Intelligence in Machines: Computing Contextual Appropriateness of Affective States, to appear in Proceedings of IJCAI 2009 - Twenty First International Joint Conference on Artificial Intelligence, Pasadena, USA (2005)

[Asimov 50] Asimov, I.: I,Robot, Spectra (2004)

[Moor 79] J.H. Moor,: Are There Decisions Computers Should Never Make?, Nature and System, vol. 1, no. 4 ,pp. 217-229 (1979)

[Anderson 05a] Anderson, M., Anderson, S.L., Armen, C.: Towards Machine Ethics: Implementing Two Action-Based Ethical Theories, Proc. AAAI 2005 Fall Symp. Machine Ethics, pp. 1-16 (2005)

[Anderson 05b] Anderson, M., Anderson, S.L., Armen, C.: MedEthEx: Toward a Medical Ethics Advisor, Proc. AAAI 2005 Fall Symp. Caring Machines: AI in Eldercare, AAAI Press, pp. 9-16 (2005)

[McLaren 06] McLaren, B.M: Computational Models of Ethical Reasoning: Challenges, Initial Steps, and Future Directions IEEE Intelligent Systems, vol. 21, no. 4, pp. 29-37, July/Aug. (2006)

[Guarini 06] Guariani, M.: Particularism and the Classification and Reclassification of Moral Cases, IEEE Intelligent Systems, vol. 21, no. 4, pp. 22-28, July/Aug. (2006)

[Arkoudas 05] Arkoudas, K., Bringsjord, S.: Toward Ethical Robots Via Mechanized Deontic Logic, Machine Ethics: Papers from the AAAI Fall Symp., AAAI Press (2005)

[van den Hoven 02] van den Hoven, J., Lokhorst, G.J.: Deontic Logic and Computer-Supported Computer Ethics,Cyberphilosophy: The Intersection of Computing and Philosophy, J.H. Moor and T.W. Bynum, eds., Blackwell, pp. 280-289 (2002)

[Wiegel 05] Wiegel, V., van den Hoven, J., Lokhorst, G.J.: Privacy, Deontic Epistemic Action Logic and Software Agents,Ethics of New Information Technology, Proc. 6th Int'l Conf. Computer Ethics: Philosophical Enquiry (CEPE 05), Center for Telematics and Information Technology, Univ. of Twente, pp. 419-434 (2005)

[Wallach 09] Wallach, W., Allen, C.: Moral Machines - Teaching Robots Right from Wrong, Oxford University Press (2009)

[Grau 06] Grau, C.: There Is No "I" in "Robot": Robots and Utilitarianism, IEEE Intelligent Systems, vol. 21, no. 4, pp. 52-55, July/Aug. (2006)

[Moor 06] J.H. Moor,: The Nature, Importance, and Difficulty of Machine Ethics, IEEE Intelligent Systems, vol. 21, no. 4, pp. 18-21, July/Aug. (2006)

[Aristotle 24] Aristotle, Nicomachean Ethics, W.D. Ross, ed.,Oxford Univ. Press (1924)

[Kant 1975] Kant, I.: Groundwork of the Metaphysic of Morals, Practical Philosophy, translated by M.J. Gregor, Cambridge Univ. Press (1996)

[Mill 1871] Mill, J.S.: Utilitarianism, fourth edition. Longmans, Green, Reader, and Dyer, London (1871)

[Rzepka 08] Rzepka, R., Higuchi, S., Ptaszynski, M., Araki, K.: Straight Thinking Straight from the Net - On the Web-based Intelligent Talking Toy Development, The Proceedings of the IEEE International Conference on Systems, Man and Cybernetics (SMC'08), pp. 2172-2176, Singapore (2008)

[Nakamura 93] Nakamura, A.: Kanjo hyogen jiten [Dictionary of Emotive Expressions] (in Japanese), Tokyodo (1993)

[Bentham 1789] Bentham, J.: Introduction to the Principles of Morals and Legislation, W. Harrison, ed., Hafner Press (1948)

[McDermott 08] McDermott, D.: Why Ethics is a High Hurdle for AI, presented at North American Conference on Computers and Philosophy (NA-CAP), Bloomington, Indiana (2008).

[Powers 06] Powers, T.M.: Prospects for a Kantian Machine, IEEE Intelligent Systems, vol. 21, no. 4, pp. 46-51, July/Aug. (2006)