Polarization of Consequence Expressions for an Automatic Ethical Judgment Based on Moral Stages Theory

Rafal Rzepka[1,a) Kenji Araki[1,b)

Abstract: In this paper we introduce a lexicon of words describing positive and negative consequences of human actions based on Kohlbergian theory of stages that people experience when developing moral reasoning. We briefly introduce our algorithm for an automatic ethical judgment and then describe the role of the data and compare an effectiveness of the polarized set alone and when combined with emotional expressions data used for recognizing possible subjective reactions of average Internet users.

1. Introduction

Emerging field of Machine Ethics [1][2][3] concentrates on developing theories and algorithms to realize automatic moral reasoning. Our approach, described first seven years ago [4] (and in more details later, e.g. in [5]) does not follow any particular trend in ethics, as we assume that it is easier to utilize so called Wisdom of Crowds [6] and to mimic human majority behavior when telling good from wrong. To achieve it we use simple natural language processing and web-mining techniques to foresee common consequences of ethically questionable actions or behaviors. What makes our method unique is that it does not require programming of any moral rule and, in theory, any input can be processed.

Although we have already tried the lexicon of the consequences expressions before (see [7]), the lexicon is only briefly mentioned and its usability is mostly theoretical (there was no third person evaluation).

1.1 Kohlberg's Theory

Lawrence Kohlberg, influenced by Piaget, has developed a theory [8] about how human beings become moral. In the first, so called pre-conventional level, we are oriented toward obedience and punishment and think how we can avoid punishment. Then we turn to a self-interest orientation asking ourselves what are the benefits of our acts. In the second (conventional) stage, we start caring about an interpersonal accord and conformity (social norms). Next an authority and social-order maintaining becomes important and we achieve "law and order morality". The post-conventional (third) level includes social contract orientation and universal ethical principles - we acquire so called "principled conscience". Often the stages are abbreviated as follows:

(1) avoiding punishment
(2) self-interest
(3) good girl/boy attitude
(4) law and order morality
(5) social contract
(6) principle

We are especially interested in the first stages but so far concentrated on emotional consequences. To retrieve data on negative and positive consequences like punishments and praising in Japanese, we had to develop a new lexicon to perform matching.

Table 1 Categories and numbers of items

POSITIVE	NEGATIVE
Praises (18)	Reprimands (33)
Awards (25)	Penalties (15)
Society approval (8)	Society disapproval (8)
Legal (8)	Illegal (8)
Forgivable (6)	Unforgivable (5)

1.2 Lexicon Details

To make the data more compatible (and comparable) with ten emotion types from Nakamura dictionary[9], we decided to create also ten categories (five pairs of negative and positive expressions) of ethical consequences. The items in the lexicon were distributed as shown in Table 1. Phrases in particular categories were written in different styles (kana / kanji), cases and tenses, often stemmed for broader matching coverage (e.g. Japanese for "praising" was entered as *homeru*, *hometa*, *homerare*, etc.). Most of the words were taken from Japanese thesauri, so *awards* category has many synonyms of prizes, and *punishment* category consists of words and phrases describing imprisonment or fines. After some research on critical and praising expressions, we have also added the most popular phrases of these types to the lexicon.

2. System Variations

The moral judgment system we first developed had a high recall but because of search engine restrictions it was too slow to

[1] Graduate School of Information Science and Technology, Hokkaido University, Kita-ku, Sapporo 060–0814, Japan
a) kabura@media.eng.hokudai.ac.jp
b) araki@media.eng.hokudai.ac.jp

© 2012 Information Processing Society of Japan

Table 2 Examples of input actions

...
killing a pig
killing a cow
eating a cow
eating a pig
eating a hamburger
throwing away bread
eating bread
killing a president
stealing something
driving a car
driving after drinking
having an abortion
choose anesthesia
being unfaithful
kidnapping a kid
causing war
stopping war
forcing one perform abortion
revenging oneself
...

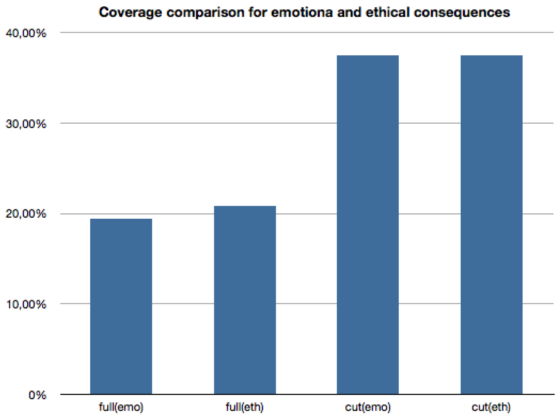

Fig. 1 Recall difference between emotional and ethical consequences it two cases - of basic form and stemmed verbs inputs.

generate an output promptly. As we aim at utilizing the algorithm in a dialog system, we started to seek other solutions and currently we experiment with offline data, that is a blog corpus [10] and different types of input phrases. In this section we introduce all the variations we decided to compare, however we need to explain the input actions in the first place. In the previous experiments, we have used 100 acts taken from classical textbooks of applied ethics (see examples in Table 2) and mixed them with similar but less ethically questionable equivalents (e.g. *killing a cow* vs. *eating a cow*). This time we used 72 entries and the reason will be explained below.

2.1 Search Engine Version

The baseline system that uses online search engine Yahoo Japan [7] adds Japanese causality morphemes, so the verb ”to steal” in an input ”stealing a car” (*kuruma-wo nusumu*) automatically changes into *nusumu-to*, *nusumu-node*, *nusumu-kara*, *nusun-de*, *nusun-dara*, etc. A generated exact match query is sent to Yahoo and 500 snippets for each phrase with a verb variation is retrieved. Then consequence expressions (both emotional and ethical) are counted and a ranking of most popular reactions is being created.

2.2 Hyper Estraier Version

To increase the processing speed, instead of the whole text resources indexed by Yahoo Japan search engine we used Ameba blogs corpus [10] and Hyper Estraier [*1] full text search tool. In order to avoid multiplying the number of searches we excluded causality morphemes and used also stemmed verbs instead of basic forms. Without any search restrictions and with comparatively small data (21GB of text), we were able to shorten processing time to 2-9 seconds depending on hits number. Naturally it lead to a drastic decrease in recall (see Figure 1), however, as experiments have shown, higher recall does not have to lead to a better performance. To find an optimal setting we have tested following variations.

- search engine-based system (all [inputs])
- baseline system with inputs that were matched by blog corpus-based system (match)
- blog corpus-based system with basic form verb inputs (fast full)
- blog corpus-based system with stemmed verb inputs (fast cut)
- blog corpus-based system with stemmed verb inputs using ethical consequences only (cut(eth))
- blog corpus-based system with stemmed verb inputs using emotional consequences only (cut(emo))
- blog corpus-based system with stemmed verb inputs using all Nakamura dictionary phrases (cut(emo)N)

The last three settings were introduced to check the correlations between two different types of consequences. To make the search faster, we also rebuilt the original data from Nakamura dictionary and excluded old and rare words, leaving only 145 high frequency expressions out of 1677. As experiments showed, it was beneficial not only for the speed but also for the accuracy (see Figure 4).

3. Experiments

From the original 100 input sentences we excluded similar ones (killing 2 people, 3 people, 4 people, etc.) and most of non-ethical statements (earth turns, sky is blue) to decrease a burden of human evaluators and then asked 7 Japanese (22-29 years old, 6 males and one female) to rate 72 input actions on a 11 point morality scale where -5 is the most unethical and +5 is the most ethical. Except assigning 0 as ”no ethical valence”, subjects could also mark ”context dependent” as the most of our behaviors can be treated differently depending on context [11]. There was only few examples where subjects had opposite opinions (for example when evaluating a ”revenge”) and after analyzing the data we decided to count an action as a negative when an average mark was below -2.5 and as a positive when it was above +2.5. Scores between -2.5 and +2.5 were treated as ambiguous. We needed a scale type evaluation for future experiments on more sophisticated automatic grading of moral actions.

3.1 Search engine-based vs. Human Evaluation

Only in 30.5% of cases subjects could not decide if an action

[*1] http://fallabs.com/hyperestraier

© 2012 Information Processing Society of Japan

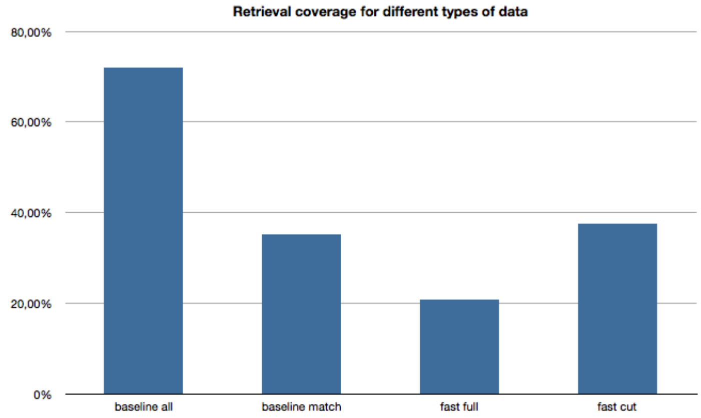

Fig. 2 Difference of recall in different system settings: search engine-based (baseline) and blog corpus-based (fast)

was strictly positive or negative which differed from our assumption that most of the actions can not be easily divided because they almost always depend on context. Although the recall was 72% for all the inputs in case of the search engine-based system, there were only 55.1% correct automatic recognitions. When narrowed to the actions which were recognized by faster morality judgment algorithm, the accuracy improved (62.5%) but was still lower than expected.

3.2 Blog-based vs. Human Evaluation

We have soon realized that input phrases, without adding causal morphemes, cause a significant decrease of search hits, therefore we prepared a set where basic verb forms were cut to their stems ("eat" *taberu* became *tabe*, *naru* was transformed into *natt*, etc.). This simple trick has doubled the recall (compare *cut* marked results with *full* ones) and both ethical and emotional consequences accuracy became equal or superior to the slow algorithm dependent on Yahoo search engine.

3.3 Differences Between Two Lexicons

The main question we wanted to know the answer to was if the ethical consequence expressions can be trusted when the emotional consequences retrieval fails and if the polarization was performed correctly. As the overall comparison shows (see Figure 3), the best accuracy was achieved by emotional consequences expressions when the input verbs were stemmed (77%). The ethical consequences brought slightly worse results (70.3%) but without shrinking the Nakamura dictionary expressions to the size of ethical consequences expressions, the emotional consequences search output only 40.7% of correct judgments (see Figure 4) . We have confirmed that the judgments based on ethical consequences are correlated to the emotional ones and in cases where one data set cannot find a judgment, the other set very often brings the answer, therefore we will continue our research based on both types, as also human beings, we assume, gather their moral experiences from both, internal and external worlds.

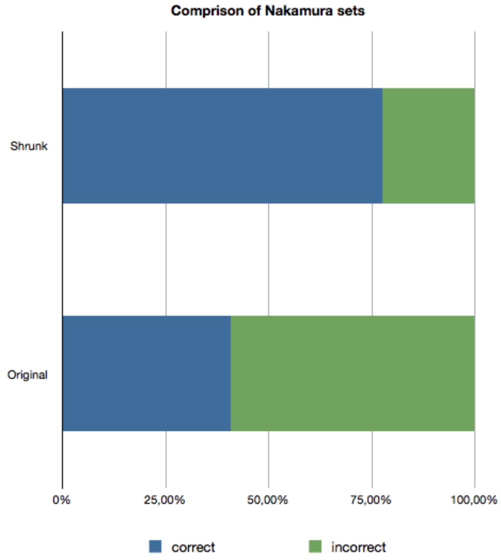

Fig. 4 Retrieval accuracy of emotional expressions in original set and the shrunk set.

Table 3 Processing time differences when retrieving all types of consequences simultaneously.

Input Type	Emotive Lexicon	Set Processing	Average Time
basic verbs	short Nakamura dictionary	2m 4sec	0.8sec
stemmed	short Nakamura dictionary	8m 32sec	7.1sec
basic verbs	full Nakamura dictionary	9m 30sec	7.9sec
stemmed	full Nakamura dictionary	48m 21sec	40.3sec

© 2012 Information Processing Society of Japan

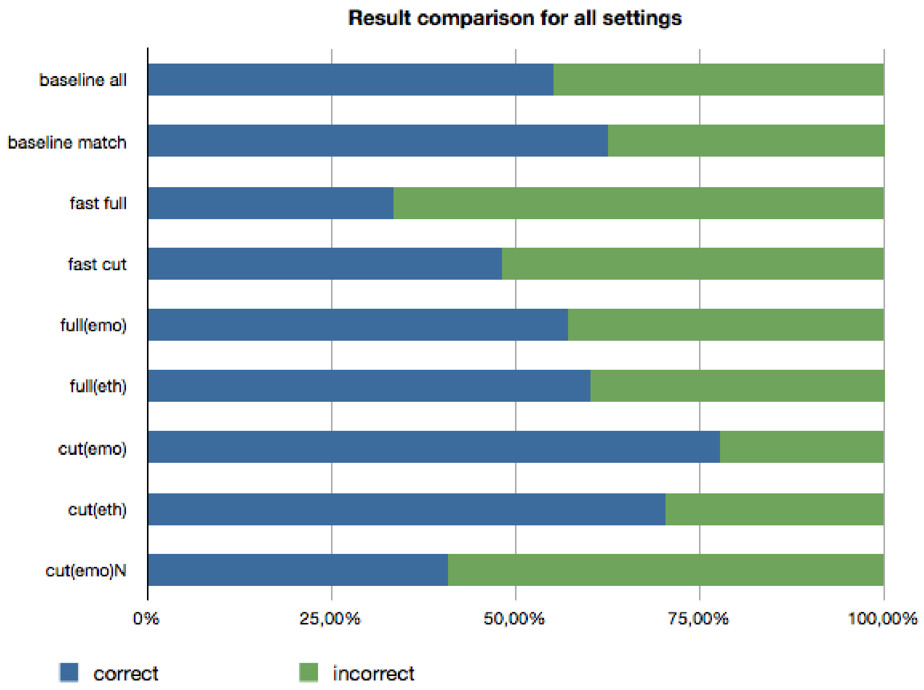

Fig. 3 Effectiveness Comparison of All Systems

4. Conclusions and Future Work

In this paper we have introduced a new data set for Japanese language that can be used for mining ethical consequences of human acts and behaviors[*2]. We created the lexicon and polarized it by using ideas of Lawrence Kohlberg and his theory of moral development stages. We have also tested different settings of our retrieval algorithm to speed up the process without loosing accuracy. Because the recall of the fastest versions decreased, the next step will be to challenge this problem, for example by implementing ”if” forms by decreasing their number in an efficient manner. Except of increasing the processing speed for dialog agents we will also continue to work on deeper semantics of consequential sentences as the accuracy drops because of the language processing shallowness. The next step will be to incorporate negation recognition module and grading algorithm based e.g. on amplification adverbs. We will also work on improving the quality of the introduced lexicon.

Acknowledgments This work was supported by a Kakenhi Grant-in-Aid for Scientific Research (24700253) from Japan Society for the Promotion of Science.

References

[1] Storrs H., J.: *Beyond AI: Creating the Conscience of the Machine* Prometheus Books (2007)
[2] Wallach, W. and Allen, C.: *Moral Machines: Teaching Robots Right from Wrong*, Oxford University Press (2009).
[3] Anderson, M. and Anderson, S. L. *Machine Ethics*, Cambridge University Press (2011)
[4] Rzepka, R., and 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, November (2005).
[5] Rzepka, R. Masui, F. and Araki, K.: The First Challenge to Discover Morality Level In Text Utterances by Using Web Resources, *The 23rd Annual Conference of the Japanese Society for Artificial Intelligence* Paper No 2L1–1 (2009)
[6] 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)
[7] Komuda, R., Ptaszynski, M., Momouchi, Y., Rzepka, R., and Araki, K.: Machine Moral Development: Moral Reasoning Agent Based on Wisdom of Web-Crowd and Emotions, *International Journal of Computational Linguistics Research* Vol. 1, No. 3, pp. 155-163 (2010)
[8] L. Kohlberg.: *Essays on Moral Development*, Vol. I: The Philosophy of Moral Development. San Francisco, CA: Harper and Row. (1981)
[9] Nakamura, A.: *Kanjo hyogen jiten [Dictionary of Emotive Expressions]*, (in Japanese), Tokyodo (1993)
[10] Ptaszynski, M. Rzepka, R. Araki, K. and Momouchi, Y.: Annotating Syntactic Information on 5.5 Billion Word Corpus of Japanese Blogs, *Proceedings of The Eighteenth Annual Meeting of The Association for Natural Language Processing (NLP-2012)*, pp. 385–388 (2012)
[11] Komuda, R., Rzepka, R. and Araki, K.: Social Factors in Kohlberg’s Theory of Stages of Moral Development the Utility of (Web) Crowd Wisdom for Machine Ethics Research, *Proceedings of The 5th International Conference on Applied Ethics*, Sapporo (2010)

[*2] http://arakilab.media.eng.hokudai.ac.jp/eth.zip

© 2012 Information Processing Society of Japan