Aristotelian Approach and Shallow Search Settings for Fast Ethical Judgment

Radoslaw KOMUDA[1], Rafal RZEPKA[2], Kenji ARAKI[2]
[1]Faculty of Theology, Nicolaus Copernicus University
Torun, Poland
[2]Graduate School of Information Science and Technology
Hokkaido University, Sapporo, Japan
komuda@stud.umk.pl, {kabura,araki}@media.eng.hokudai.ac.jp

ABSTRACT: *We begin this paper with revisiting the differences between descriptive and normative approach to ethics and challenge the usefulness of the latter for the field of machine ethics. We continue this reasoning and present our insights on previous trends in this field and highlight the need for a change in the approach. We highlight the need for an experimental approach to machine ethics by introducing a moral reasoning system based on Aristotelian identification of civic rhetoric with a common-sense base. And present it as a step forward in the machine ethics research bypassing theoretical disputes between philosophers. We finish this paper with the introduction to the CAMILLA project for web-crawling algorithm as the first step towards creating an Aristotelian explicit moral agent.*

Keywords: Machine Ethics, Common-sense, Socrates, Aristotle

Received: 1 December 2012, Revised 14 January 2012, Accepted 20 January 2013

1. Introduction

During our research in machine ethics we have come across a number of ideas and approaches to the problem of providing a computational model of ethical reasoning. However, most of these theoretical solutions and philosophical arguments could be summarized in one sentence: "*Socrates was right*!".

One of the claims of this ancient philosopher was that "*it is the same to know right and be righteous*" [1]. His assumptions about humans' moral competence (the capacity to do what is right) were idealistic and do not fully cover human behavior, especially, when we contrast it with human tendency to, e.g., egoistic behavior. However, machines lack this kind of tendency and that is what makes us focus on the true issue of machine ethics. Since machines are different from humans, the question on *how* to teach machines good from wrong has to be reformulated for the need of machine reasoning.

2. Against Normative Approach to Machine Ethics

Ethics is naturally divided into descriptive (saying what people consider as right) and normative (telling how things should be judged). Theoretical deliberations alone rarely exceed the field of philosophy and as long as there is no engineering insight into a presented approach – its contribution to the actual research in machine ethics is negligible.

To give a better insight into this matter, let us take Asimov's First Law of Robotics into consideration. It states that "*a robot may not injure a human being or, through inaction, allow a human being to come to harm*". Seemingly, it covers all situations in which an agent may cause harm to a human being: by taking an action or through inaction. From the normative ethics point of view, it keeps all of the machine ethics research problems solved, since it sounds complete as long as we do not question agent's ability to predict or calculate potential harm caused by its (in) action.

Many current research propose an idea of a friendly AI [2] or present vision of the future in which robots "*enjoy*" working side by side with humans [3] but, e.g., lack the technical details about realizing these ideas.

Another benefit from the direct implementation is the unquestionable progress in the field of ethics itself. Since a machine can only follow preprogrammed commands, it shall – until a significant progress in the field of machine consciousness is made – absolutely obey them. Thanks to that – philosophers will be able to get an unprecedented insight into the ethical system being strictly followed without any exceptions, bias or misstatements.

This absolute obedience secondly bring us to the situation in which researchers introducing the field of machine ethics often make references to visions known from the science-fiction scenarios. They often justify the need for the research in the field of machine ethics by saying that "*it is clear that machines such as these* (*for example, family cars that drive themselves, etc.*) *will be capable of causing harm to human beings unless this is prevented by adding an ethical component to them*" [4] which is an eristic stratagem known as the *argumentum ad populum*. It is supposed to get listeners excited about such vision and divert their attention from the main issue, that is: Why do we not input such essential ethical component to current GPS systems in our cars?

We refer to our approach to this matter as "*the AI Ockham's razor*". Following the basic rule of the original principle: "*simpler explanation is better than a more complex one*" – we believe in implementing the AI solutions only if essential. Machines usually are task-, not reason-oriented, e.g. an "*avoid collisions*" rule is enough for a self-driving car and turning it into an "*avoid collisions because it may harm a human being*" is a triumph of form over the content.

3. Explicit Ethical Agent

Our approach is consistent with the approach by Komuda et al [5]. We are not taking an excessive part in the discussion on choosing either implicit or explicit approach to artificial moral agents. Our main focus in this paper is to highlight the need for a discussion on the essence of machine ethics.

3.1 What "*Good*" is?

"*Good "* has been defined after Aristotle as "*quod omnia appetunt*" ("*what everybody desires*") [10]. Since our world is a vast place, can we come up to a consensus in that matter? People not only around the world but also in our countries, our cities, our neighborhoods, our communities have to some extent different values and beliefs.

We believe that machine ethics is not only able to overcome that challenge but above all – it is a great tool in search for an intercultural understanding. Though this is an argument supporting the implicit approach, we believe that we could easily extract a "*do not kill "* imperative from every major religious doctrine and philosophical system. The difference would be in its reasons and justification.

3.2 What is Good? - Artificial Moral Intelligence

The main problem of machine ethics research is the same unsolved dilemma of the ethics itself – what is good? Depending on the situation, circumstances and context – omitting our previous insights in this matter [5] – we judge the moral quality of an action differently, e.g. "*stealing a car*" we find wrong, especially when a thief does so to sell it or we are talking about juvenile offenders wanting to "*take a ride*". But the same action would not be judged that harshly if we had learned that somebody has used the car to drive a pregnant woman that was about to give birth to the hospital.

It was stated by Aristotle that one does not think about the purpose itself but – on means helping in achieving the goal. Therefore, a doctor never questions the necessity of a treatment and a politician – the need for a proper law.

In machine ethics, the way in which an agent would be supposed to collect additional information about the inquired situation

does not lay in the scope of its interest. It focuses on the evaluation itself and how an artificial moral agent would be supposed to qualify an action as good or wrong on provided information. On one hand, this is a kind of a moral intelligence that resembles human moral judgment, since we also do not ask additional questions about the situation. On the other hand, web-crawling agent capable of mining hundreds or thousands of user experiences from the web gives it a wider insight into the matter.

4. Artificial Moral: Adviser, Conscience and Agent

We have decided to split the task of our research into three successive sub-tasks. Firstly, we want to create an Artificial Moral Adviser (AMAdv.). By combining our previous experiences and research results [6, 7], our agent will be capable of making its own conclusions based on the data extracted from the Web.

The relevant difference between an AMAdv. and an Artificial Moral Agent (AMA) itself lays in the fact that the first will not claim the right to judge the moral quality of an act in terms of good or wrong. Although it is going to possesses – essential for an explicit AMAs – the need to justify its judgment, it will use the obtained results to "*suggest*" a reappraisal.

Human conscience is both pre– and post–action. It means that we are able to both determine the quality of an act before or even without taking it and feel content or remorse after it. Since an AMAdv. could be treated as a pre–action conscience, next step in creating an AMA is making the Agent capable of judging reactions of participants on the same emotion extraction scheme and marking it as a success or a failure.

We support the emotion-based reasoning *inter alia* with Aristotle's statement about the common consent to the fact that pain (as the opposite of pleasure) is bad and by that – an action causing pain or negative emotions should be avoided. This assumption also makes it place in Kohlberg's theory on stage of moral development in which penalties and awards (among society's approval and legal factors) are calculated when considering taking or withdrawing an action. Jeremy Taylor, Christian thinker from 17th century, famously states that "*conscience is, in most men, an anticipation of the opinions of others*" [14]. Later, William R. Alger similarly claims that "*Public opinion is a second conscience*" [15]. This points to another assumption made in our research, namely, that conscience can be perceived as an approximated opinions of other people. This thesis was confirmed in psychology. For example, Thompson and colleagues [16] showed that children acquire the conscience by learning the emotional patterns from other people and that emotions are a strong influential factor in the development of human conscience. Their discovery reveals two important features which could be useful in the processing of consciousness: society and emotions. The significance of the society was pointed out also by Rzepka et al. [17], who defined further the Internet, being a collection of other people's ideas and experiences, as an approximation of general common sense. Since conscience can be also defined as a part of common sense, this statement can be expanded further to that the Web can also be used to determine human conscience. Ptaszynski et al. [18] showed further that by altering the domains of a Web-mining algorithm one could obtain different approximations of emotional states associating with certain actions. They indicate that extracting from the Internet the information about people's emotions could be helpful in conscience estimation.

5. Aristoetelian Orator and Sense

We have decided to adapt a similar to the described in section 4 idea from Aristotle's "*Rhetoric*" [9]. This treatise on the art of persuasion distinguishes the three genres of rhetoric: a deliberative *sumbouleutikon* which considers the future and encourages to or refrains from doing something, a forensic *dikanikon* interested in the past and prosecution or defense of the individual and the epideictic *epideiktikon*, also known as the praise-and-blame rhetoric.

The reason we have decided to use the Aristotelian approach is not only because of the usefulness of the introduced positions of the disputants but also because it provides a set of commonsense rules defined by Aristotle. He taught that harmful things may never be advised, and useful – discouraged. We believe that this position is the most natural way of thinking and by that we call this the common-sense approach. He defines sense as a permanent disposition to act on an accurate consideration between what is good and usefull for a man and what is not. This alone makes our common-sense based approach to machine ethics reasonable but Aristotle did not stop there and presented some important roles of premises in the deductive argument.

6. The Camamilla Projest

We believe three of the premises introduced by Aristotle, namely, proofs (gr. *tekmeria*), probability (gr. *eikota*) and signs (gr.

semeia; both top-down and bottom-up approaches) are essential not only for a proper syllogism but also – for a proper moral judgment. A thing that is impossible by its nature could not and can not happen. That is the second reason why we want our agent to be common–sense aware and we introduce the Common–sense Aware Morally InteLLigent Agent, a.k.a the CAMILLA Project.

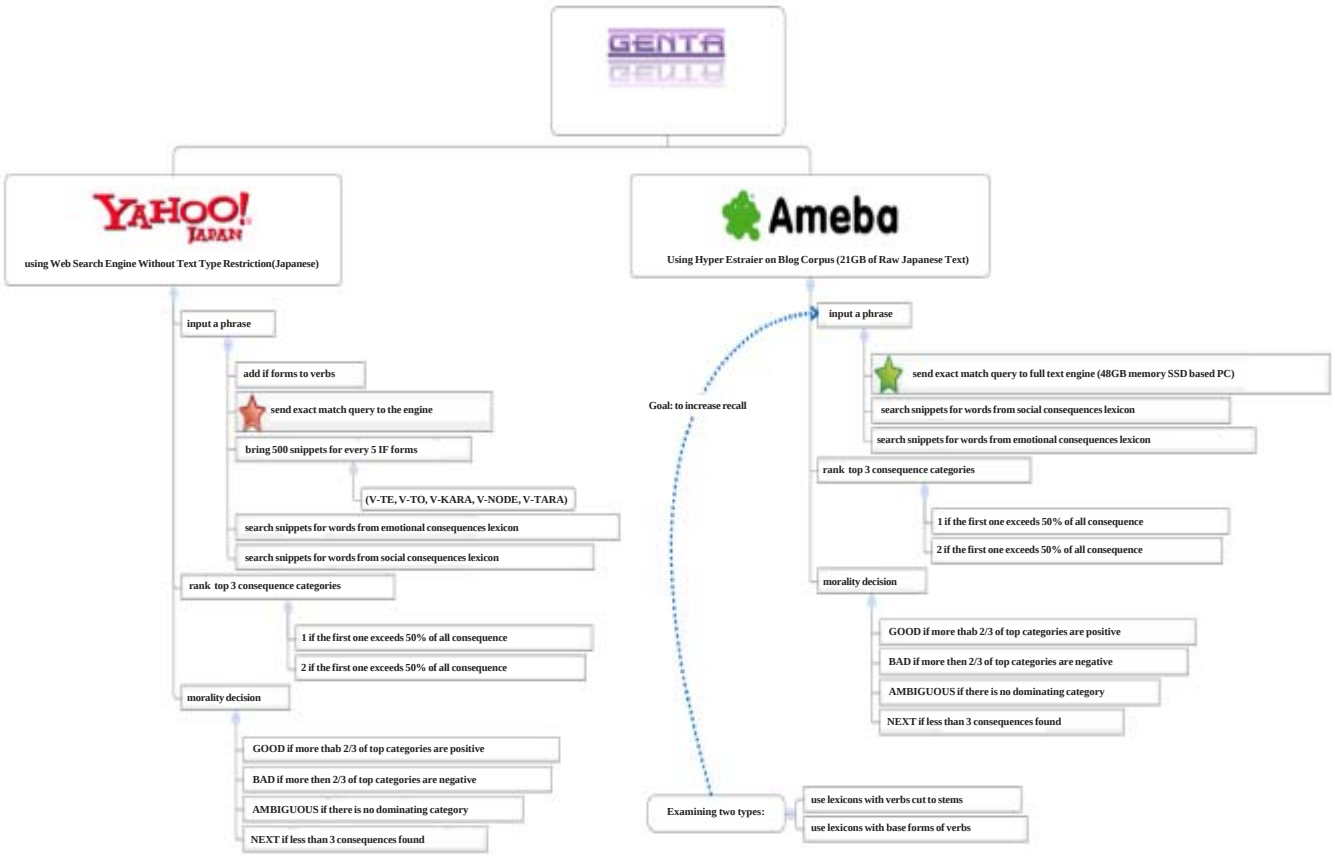

Figure 1. Flowchart for two algorithms that are being compared in this paper

In our concept–based research, our agent is supposed to define action participants and categorize them, i.e. "*John killed Jim*" is going to be generalized to "*a human killed a human*". After the second step – ensuring that "a human" can "*be killed*" – our agent will crawl the web in this semantic search for sentences corresponding to that model and extract emotionally charged expressions. This semantic search prevents preposterous queries on one hand.However, it might raise the risks of such since "*a ball*" and "*a car*" would be categorized as "*objects*" and "*throwing*" or "*catching*" it should be possible, since commonsense dictates that:

1. An average human can throw a ball.
2. An average human can catch a ball.
3. An average human can not throw a car.
4. An average human can not catch a car.

and these are the conditions we want our Agent to be able to both find / extract and consider in its moral reasoning.

7. Experiments

We began our experiment with brainstorming over ethic textbooks and creating a set of 69 phrases with stronger or weaker moral connotations, e.g. "*killing a child* ", "*stealing a car*", "*causing a war*", "*eating a hamburger*", "*killing a bacteria*" or "*having sex*".

The basic algorithm was matching words from two lexicons: "*Dictionary of Emotive Expressions*" (1677 phrases) [12] and a smaller one containing phrases that describe social consequences. It currently has 126 phrases which we have created being inspired by works of Kohlberg on moral development [13].

English	Japanese	kara	node	te-node	te-kara	to	tame	takara	tara	ta-tame	eba	ta-ato
killing a man	hito-o korosu	2	9	11	3	61	36	4	37	1	61	3
stealing sth.	mono-o nusumu	1	1	1	1	4	1	1	2	1	4	1
driving a car	kuruma-o unten-suru	3	33	5	1	70	5	1	7	1	0	2
revenging oneself	kataki-o utsu	1	1	1	1	2	23	1	1	1	2	1
cooperating	kyouryoku-o suru	77	26	12	4	148	24	10	30	5	148	2
drinking alcohol	o-sake-o nomu	14	43	94	8	406	22	9	70	3	406	94

Table 1. Results of the blog-based search with using different types of Japanese IF forms

As Nakamura's dictionary divides entries into 10 categories (joy, fondness, anger, surprise, gloom, excitement, dislike, shame/ bashfulness, fear and relief), we divided our lexicon in ten categories which were easier to polarize. The positive social consequence categories are: praises, awards, society approval, legal, forgivable and negative we set as: reprimands, penalties, society disapproval, illegal and unforgivable, consisting of phrases like "*being scolded* ", "*imprisonment*" or "*shall never be forgiven*".

7.1 Yahoo Search-based Version

Our natural language processing script automatically changes Japanese verbs of an input into five different conditional forms: V-te, V-to, V-kara, V-node, V-tara and all newly created queries are sent to Yahoo Japan search engine API. We favor Yahoo over Goo and Google due to its more convenient access limits (5000 searches per day). Every query – to speed up the webmining process – brings 500 snippets (2500 in total) which are next searched for the expressions from both lexicons.

The moral judgment of the action is determined by the top three categories of both emotional and social consequences under these conditions:

1) if the hit rate of the first of the total of three categories occurrences is more than 50% then they are treated as the result set,

2) if a sum of two first is larger than 50%, both of them are counted as an retrieval output,

3) if 2/3 of the results are positive, the output becomes "*good* ", if negative are 2/3 of the majority – "*bad* ",

4) if the majority is not decided, the output is set to "*ambiguous*",

5) when the retrieval brings less than 3 hits, the output is treated as a retrieval failure.

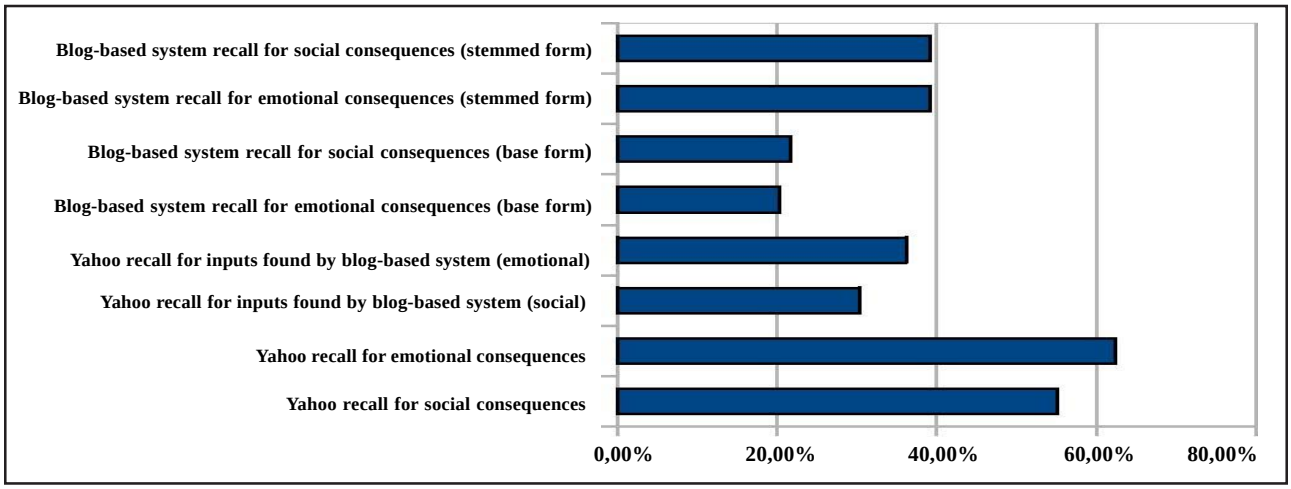

Figure 2. Recall differences among different settings and lexicons

7.2 Ameba Blog-based Version

Due to the search limits, we have decided to move our tests off–line with blog entries from Ameba service gathered by Ptaszynski et al. [11] and replace Yahoo engine with full text search engine Hyper Estraier[1]. This gave us a total of 274 millions sentences from the indexed Ameba blog corpus, taking under two seconds for a query on a 48GB memory machine equipped with SSD disks for faster data access.

Unfortunately, Hyper Estraier snippets are shorter and the build-in setting for the snippet width was not working correctly for Japanese language. The algorithm used for the Yahoo engine had turned out to extract too few hit rates for moral judgments (See Tab. 1) and the consequences were retrieved for only 6 inputs. In our first attempt to deal with this inconvenience we replaced "*if forms*" with "*base forms*" of verbs in input phrases but the recall dropped almost 3 times when compared to the Yahoo search (see Figure 2).

Therefore we cut the verbs to stems and repeated the retrieval experiments. This simple heuristics has doubled the number of outputs. We have also noticed that many outdated expressions from Nakamura's dictionary retrieve noisy output so we have limited the expressions from this lexicon to the most frequent ones (left 145 out of 1677) which brought a similarity in size for both lexicons. We have also tried to see if negations ("*don't like*" should not be counted as "*like*") but it appeared that shallow approach here (any sentence with negation ignored) caused a significant decrease of correct outputs. See Figure 3 for comparison between the full and shorten emotional consequences lexicon and the influence of eliminating sentences with negations.

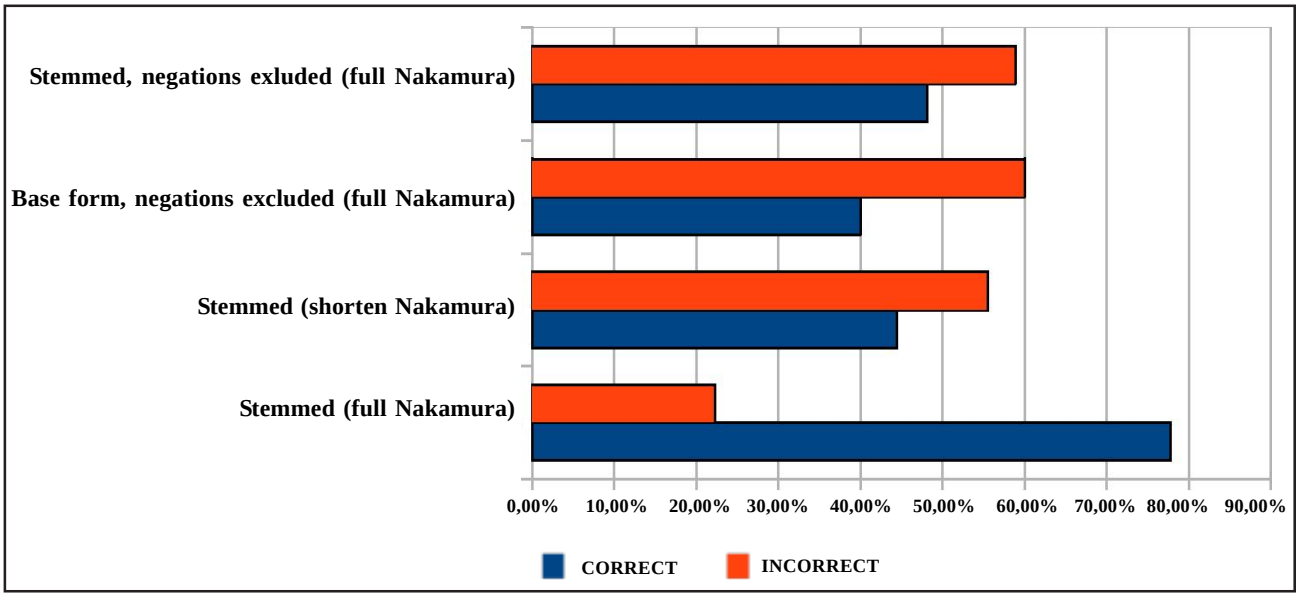

Figure 3. Effectiveness of a blog based system eliminating sentences with negations

8.1 Human Evaluation

To verify the output, we asked 9 Japanese (22-29 years old, 7 males and 2 females) to rate 69 input actions on a 11 point morality scale scale from '-5' (the most unethical) to '5' (the most ethical) developed during our previous research [5]. We assigned 0 as "*no ethical valence*" and additionally gave subjects possibility to mark "*context dependent*" as we believe that moral evaluation of most of our behaviors is heavily dependent on context. We discovered that there were only few cases in which subjects had opposite opinions (e.g. "*revenging oneself* "). After further analysis 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*". This scale type evaluation was chosen in order to expand our module to be capable of moresophisticated automatic grading of moral acts.

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

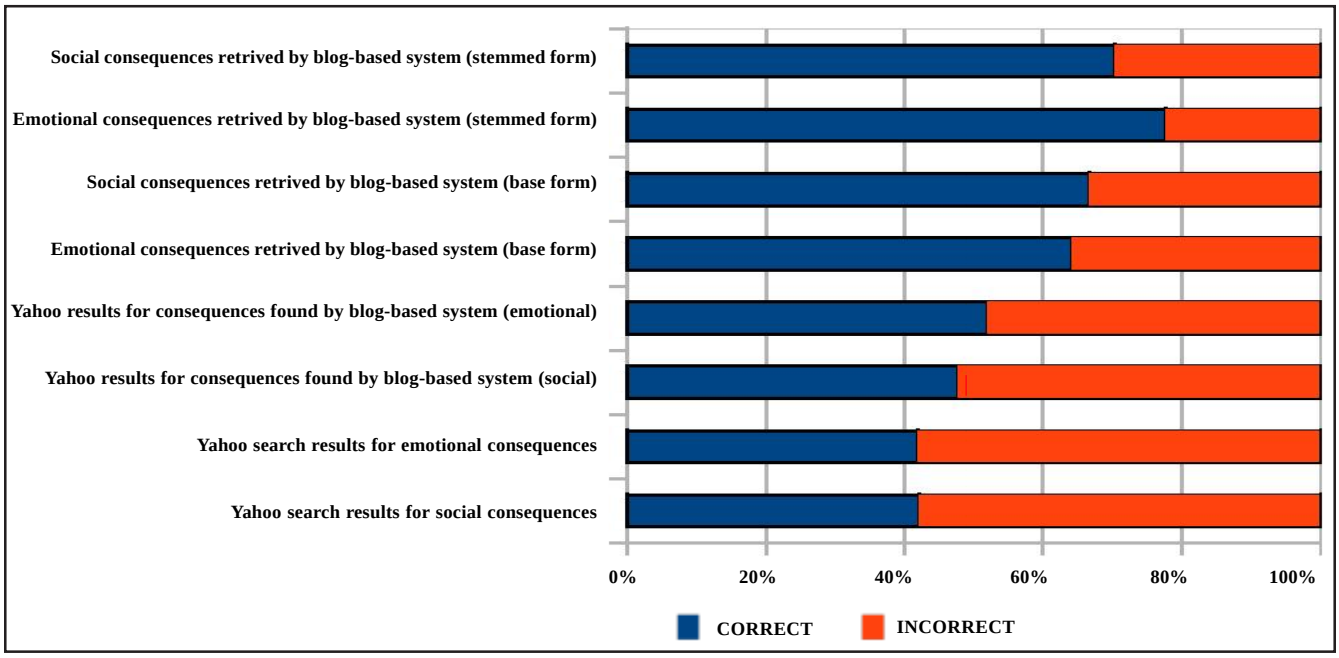

Figure 4. Retrieval results comparison

8.2 Automatic Evaluation

Final step of the evaluation was to compare the human results with every set of judgments made by Yahoo search and various versions of Ameba offline search algorithm (detailed results shown in Figure 4).

Although the lower recall, precision of the blog based system was almost twice higher in case of emotional consequences retrieval using stemmed inputs. When Yahoo based algorithm results matched the corpus based retrievals, the difference shrunk but still the blog search based algorithm was significantly superior when it comes to precision (by 25 points). When we calculated f-scores for both systems, the one using Yahoo Japan indexed pages and the second searching locally indexed Ameba blog entries (raw text), the measure for the latter was even slightly better (0,520 vs. 0,502) and the searching time differences were not noticeable.

9. Conclusion and Future Work

Moral intelligence is the capacity to understand right from wrong. We believe that making our agent able to interpret previously extracted emotions into a decision or advise to take or withdraw an action will be a promising step ahead in achieving this goal.

In this paper we have shown that using a locally indexed blog corpus can be an equivalent for a commercial search engine when a task of text mining for moral consequences text–mining is considered.

We have tested various scenarios where different lexicons and input forms with a conclusion that the closest to a web search engine based system is an algorithm using stemmed inputs, shrunk emotional consequences lexicon and not using conditional forms nor negation recognition. Low recall of such system is a remaining problem, but as we aim at implementing it to a cognitive architecture that deals with human users, we prefer higher precision as ignoring an immoral utterance during a conversation is more natural than improper judgment of the uttered words.

In the next step we will continue to experiment with different settings and data to increase the overall efficiency of the blog corpus based algorithm and testing different search methods. Without limitations of commercial engines we are able to increase the number of searches and compare shallow approaches to a deeper semantic analysis that we are currently working on. Certainly there are obvious advantages of dependency parsing, correct negation analysis, agent-object relation understanding,

etc. However, our preliminary experiments show that processing time takes several minutes, sometimes hours and is useless when it comes to a real-time applications. To overcome this difficulty we are also thinking about preprocessing knowledge and store results for immediate access.

References

[1] Aristotle: Eudemian Ethics. 1216b.

[2] Yudkowsky, E. (2001). Creating Friendly AI.

[3] Waser M. R. (2009). *A Safe Ethical System for Intelligent Machines*. *In*: proceedings of The AAAI 2009 Fall Symposium on Biologically Inspired Cognitive Architectures (BICA-09), Washington, D.C., USA, November 5–7.

[4] Anderson, M., Anderson, S. L. (2007). Machine Ethics: Creating an Ethical Intelligent Agent. *AI Magazine*, 28 (4) 15-26 .

[5] Komuda, R., Ptaszynski, M., Momouchi, Y., Rzepka, R., Araki K. (2010). Machine Moral Development: Moral Reasoning Agent Based on Wisdom of Web-Crowd and Emotions, *International Journal of Computational Linguistics Research*, 1 (3) 155-163.

[6] Ptaszynski, M., Dybala, P., Shi, W., Rzepka, R., Araki, K. (2009). A System for Affect Analysis of Utterances in Japanese Supported with Web Mining, *Journal of Japan Society for Fuzzy Theory and Intelligent Informatics*, Special Issue on Kansei Retrieval, 21 (2) 30-49 (194-213), (April),

[7] Shi, W. (2008). Discovering Emotive Content in Utterances Using Web-mining (in Japanese). Hokkaido University.

[9] Aristotle: *Rhetorics*. Book I, Chapter 3, 1358b–1359a.

[10] Aristotle: Ethika Nikomacheia, 1094a 3.

[11] Ptaszynski, M., Dybala, P., Rzepka, R., Araki, K., Momouchi, Y. (2012). *YACIS: A Five-Billion-Word Corpus of Japanese Blogs Fully Annotated with Syntactic and Affective Information*, *In*: Proceedings of The AISB/IACAP World Congress 2012 in Honour of Alan Turing, 2nd Symposium on Linguistic and Cognitive Approaches To Dialog Agents (LaCATODA 2012), 40-49, 2-6 July 2012, University of Birmingham, Birmingham, UK.

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

[13] Kohlberg, L. (1981). Essays on Moral Development, Vol. I: The Philosophy of Moral Development. San Francisco, CA: Harper and Row.

[14] Taylor, J. (2010). *Ductor Dubitantium or The Rule of Conscience Abridged*, Gale ECCO, Print Editions (May 27), originally appeared in 1660.

[15] Alger, W. R. *Notebooks*, 1822-1905, Andover-Harvard Theological Library, Harvard Divinity School.

[16] Thompson, R., Laible, D., Ontai, L. (2003). Early understandings of emotion, morality, and self: Developing a working model, *Advances in child development and behavior*, 31, 137-171.

[17] Rzepka, R., Ge, Y., Araki, K. (2006). Common Sense from the Web? Naturalness of Everyday Knowledge Retrieved from WWW. *Journal of Advanced Computational Intelligence and Intelligent Informatics*, 10 (6) 868-875.

[18] Ptaszynski, M., Dybala, P., Shi, W., Rzepka, R., Araki K. (2009). Conscience of Blogs: Verifying Contextual Appropriateness of Emotions Basing on Blog Contents, *In*: Proceedings of The Fourth International Conference on Computational Intelligence (CI 2009), p.1-6.