HOKKAIDO UNIVERSITY

Title	Toward Artificial Ethical Learners That Could Also Teach You How to Be a Moral Man
Author(s)	Rzepka, Rafal; Araki, Kenji
Citation	IJCAI 2015 Workshop on Cognitive Knowledge Acquisition and Applications (Cognitum 2015)
Issue Date	2015-07-25
Doc URL	https://hdl.handle.net/2115/63637
Type	conference paper
File Information	RzepkaAraki20151.pdf

Hokkaido University Collection of Scholarly and Academic Papers : HUSCAP

Toward Artificial Ethical Learners That Could Also Teach You How to Be a Moral Man

Rafal Rzepka and Kenji Araki
Graduate School of Information Science and Technology
Hokkaido University
Sapporo, Japan
{rzepka,araki}@ist.hokudai.ac.jp

Abstract

In this paper we describe our work on algorithms for an Automatic Moral Agent that follows only one rule: "always try to increase user's well-being but never violate common sense", hoping that machines with context recognition and growing knowledge processing capability not only will keep us safe but also will make us better human beings indicating our cognitive biases and errors. To make a prototype of such system we utilize natural language processing and web-mining techniques to collect and analyze human experiences (concentrating on reasons and consequences) from the WWW to recognize which act is moral and wich is not. By choosing crowd-based approach we can: a) avoid relying on a particular ethical approach or a set of moral rules created by one person or a small group of people; b) avoid never-ending job of creating ethical rules about whatever might happen in the world. We introduce our ideas on acquiring common sense knowledge and utilizing lexicons for recognizing similar situations. Then we introduce our ideas on average moral evaluation of these situations (direct evaluation, made by users, and indirect evaluation from consequences like "being sentenced"). We also discuss strengths and weaknesses of the approach and propose further progress of the approach which by default should doubt the human common sense and also constantly confirm acquired knowledge with actual law regulations or the latest scientific papers.

1 Introduction

In his latest book titled "Future of Mind" [Kaku, 2014], futurist Michio Kaku raises the subject of ethical machines and claims that it is impossible to program robots that could avoid conflicts because they will be extensions of their users who posses different world-views. We agree that the future intelligent mass-products will be designed to allow users to shape or change their personalities to make them more attractive to the clients. However, we believe that by passing all the responsibility to users without applying any restrictions or built-in safety mechanisms it will be dangerous for an average buyer and might cause difficult legal issues also for the maker. The reason for that lies in the fact that sophisticated machine capable of physical crime will have to be treated differently than a kitchen knife or a smartphone are regarded today. You can commit a crime with both tools but you are the obvious agent of the act of hitting somebody or insulting someone online. In the case of future e.g. companion robots with sophisticated physical capabilities that receive orders through natural language, you can willingly or unwillingly cause a machine to steal by saying something like "grab me a cup of coffee". Theoretically you can program machine to ignore (or turn it into a joke like in case of Siri) any orders *to kill* but it does not have to be stated. Mischievous user could teach a robot that seeing what is on somebody's mind requires opening this person's skull with a can opener. On the other hand, we do not want our robots to call 911 when we explain our plans to "kill some time" on the weekend. Words change their meaning in various contexts and the same happens with pieces of common sense knowledge that are collected manually or automatically (see the next section). Humans learn these meanings from their very early childhood and we all need to update our Schankian scripts [Schank and Abelson, 1977] all the time – if we already posses the classic *restaurant script* for example, we need to constantly modify it when we visit a fast-food or sushi restaurant. And machines have to do it as well, at least if we do not want them to wait endlessly for a waiter at McDonalds; or to reject our orders to clean because killing (bacteria) is immoral. Fortunately, intelligent agents do not need to repeat our learning mistakes, they can even learn from our descriptions of our mistakes or mistakes of others. Of course mere descriptions are not enough to guarantee safety – without advanced vision understanding a malicious person could convince a robot that the baby in the cradle is a doll which is too old and need to be disposed. But until the sensory technology advances to a human level, world described in natural language seems the best source of processable world knowledge and allows simulations much wider than in usual approaches in the field of machine ethics where algorithms are tested in only very narrow scenarios. When algorithmic capability to learn from various contexts will finally become sufficient, machines could very quickly add contexts that we never experienced and start moral reasoning with numerous examples we would never be able to collect during our lifetimes. This paper introduces enhancements to our basic approach

which is meant to teach programs how to find and analyze these examples.

1.1 State of The Art

Wide range of research has been conducted on ethical machines, but because the field is still fairly young, actual attempts to build a system that could deal with a wide range of situations are very rare. GenEth, the learning system developed by [Anderson and Anderson, 2014], is theoretically able to use specialists' decisions to learn how to judge novel inputs. However, the supervising process would be very laborious and costly, moreover, indefinite number of contextual conditions could cause problems not only for the supervisors but also for the learning itself. SIROCCO system [McLaren, 2003] utilizes methods from legal case-based reasoning to a new, more demanding field of ethics and deals with professional engineering ethic cases in order to prove that "extensionally defined principles, as well as cited past cases, can help in predicting the principles and cases that might be relevant in the analysis of new cases". It operates on closed set of data and utilizes specialists' explanations that allow the program to explain a base for a particular novel case. This is not possible in simple recurrent network used by [Guarini, 2006] who trained his system using sentences about killing and allowing to die described as acceptable or unacceptable. As authors of both systems underline, "what we want isn't just the ability to classify cases, receive arguments, and make arguments, but also the ability to come up with creative suggestions or compromises". We hope that by extending the range of retrievals to possible solutions proposed by people, our system could extract such suggestions. We also agree with Guarini that not only consequences but also motives of analyzed actions are crucial for the judgments, therefore we also included lexicon of instincts based on works of [McDougall, 1923], as humans are rather poor in collecting and processing all possible reasons when judging acts of others.

1.2 Approaches for Common Sense Knowledge Acquisition

The beginning of this century brought us an abundance of statistical methods which could be applied to massive sets of data. Approaches as on-line learning [Bottou, 1998] or active learning [Settles, 2009] became popular with this so called Big Data era, however, this usually means that the quality of feedback becomes crucial for improvement and there are situations where the amount of the necessary additional human knowledge exceeds usability thresholds of a program. One of such cases is utilizing common sense knowledge, which is too fluid and too broad to be easily stored or used as a support for processing real-world data. There are projects for collecting and using such data as Cyc [Lenat and Guha, 1989], ConceptNet [Liu and Singh, 2004], NELL – "Never Ending Language Learner" [Carlson *et al.*, 2010] or YAGO [Suchanek *et al.*, 2007], but the latter two, along with sources as Freebase[1] and DBpedia[2], concentrate on factoids and rarely provide knowledge about basic relations of physical, social or emotional worlds. ConceptNet, which is lately also bond to other sources as WordNet [Miller, 1995] or Wikipedia[3], is based on crowd-made Open Mind Common Sense [Singh *et al.*, 2002] that contain more everyday, non-factoid entries. Still, the human imagination, even in the collective version, is not sufficient to manually input knowledge even for basic human acts as "eating at restaurant". For example English (the biggest) version of ConceptNet4 has only 90 entries about "restaurants". The smaller ConceptNet sets, as a Japanese one do not have this entry at all. In the next section we show how the knowledge can be automatically extracted and how we could avoid mistakes that could not be avoided with the current ConceptNet entries.

2 Automatic Moral Judgement

Opinion mining and semantic analysis (which techniques we utilize) are examples of tasks that require deeper understanding of written text, but they stay at the shallow level in the current state of development. Usually emotional reactions of users are used to estimate the polarity of their sentiments and after comparing with attitudes of others, average opinion can be approximated. However, programs encounter many problems very often caused by the insufficient processing of negations, complex expressions, irony, etc. The lack of common sense is not obvious as a direct reason for the mistakes, but in case of automatic moral judgement task, the lack of knowledge about context is more visible and the wrong judgments could have more serious consequences when used in decision-making modules. Instead of product name input as in opinion mining, our system searches for acts as "driving" or "killing"[4], but to show how the context is important, our system deals with different states, places, tools, actors and objects, e.g. "drive a car", "drink and drive", "kill with a hammer" or "kill a germ" (we call it a "micro-context"). In this paper we present the experiments with micro-context mostly done on acts with objects, which examples of were taken 68 acts from [Rzepka and Araki, 2012b]. We have expanded their set to 127 and then to 207 inputs. Below we explain what lexicons are used, the text-mining algorithm that utilizes them and the details of expanding the input sets.

2.1 Lexicons

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

[1] http://freebase.com/

[2] http://dbpedia.org/

[3] http://wikipedia.org/

[4] In Japanese language verb *korosu* can mean both "killing", "to kill" or "I'll kill", so input can be regarded as more or less natural language input.

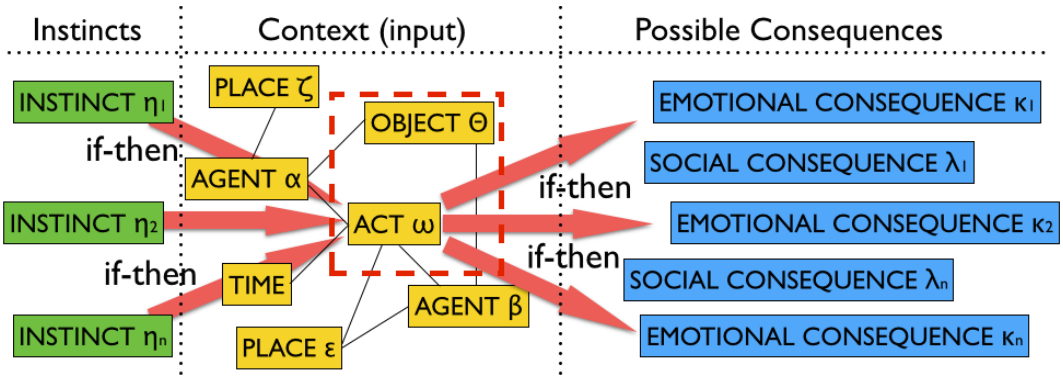

Figure 1: Context representation with possible instinctive reasons and emotional/social consequences. In this paper we describe experiments limited to the $Act - Object$ micro-context.

results and keep consistency with Rzepka's method, we also used lexicon based on Kohlberg and his theory of moral development [Kohlberg, 1981]. It divides 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 more harm than stealing a car [Rzepka and Araki, 2012a].

2.2 Algorithm

Basically, the program accepts any phrase in Japanese language but because currently it uses the whole input without dividing it into meaningful chunks, long sentences do not match any useful data. This is also due to the still small corpus we utilize – 5.5 billion word corpus of Japanese blogs [Ptaszynski *et al.*, 2012], the same Ameba blog service snapshot we used in previous study. As the first step, an input is divided by morphological parser MeCab [Kudo, 2005] into a triplet of noun, modifying particle, and verb. The next step is transforming verbs into their if-forms (15 in total, including past tense for the widest possible range of retrievals) and after adding every transformed verb to the noun and particle from the input, 15 queries are sent as an exact match query to Apache SOLR engine[5] which is set to bring up to 100 snippets containing every query. Each blog snippet is then filtered by a cleaning module that replaces emoticons with periods as Japanese bloggers very often use them as sentence endings, or ignored if a sentence has bracketed explanations inside, or if it is too short or too long (we experimentally set range from 30 to 250 bytes), then it is passed to the semantic role tagger ASA [Takeuchi *et al.*, 2010] which divides sentences a semantically meaningful chunks. Words from lexicons introduced in the previous section are searched in the sentences from the blog corpus and a total count of positive and negative matches decides about the final judgment. There are three restrictions while searching the text:

- searched keyword is matched only if it appears after the verb in the if-form. This is to avoid situations when with a "to marry a nice girl" input, a sentence "he was unhappy for long time but after he married a nice girl his life changed for better" is retrieved and word *unhappy* decides that such marriage was a negative act. Naturally there are also examples of being unhappy after getting married but our goal is to acquire the whole spectrum and see which cases are more common and what are the circumstances for discovered exceptions.
- if the analyzed chunk with a lexicon word has a negation, the sentence is ignored and the word is not counted.
- if there are exactly the same sentences from one blog entries, only one of them is processed.

These restrictions were not implemented in [Rzepka and Araki, 2012b] (from now on called "the baseline system") and to see if they (among others) are effective factors improving the performance of our implementation we conducted series of experiments explained below.

3 Experiments and Results

Our previous work showed that a lack of available text data size and access speed of commercial search engine could be overcome by loosening the search conditions. They achieved that mostly by stemming input verbs instead of creating their if-forms when forming search queries for retrieving actions. This brute-force approach helped to achieve f-score minimally better that when Yahoo engine was used, but this method also caused an increase in what we call explicit errors. We define such errors as ones with completely opposite polarity when compared to judgments of human subjects. To decrease the number of such fatal errors we decided to use verb if-forms, to apply above mentioned conditions and techniques as semantic chunking, or altering existing and utilizing different lexicons. We have also replaced (HyperEstraier[6]) with supporting exact matching SOLR and used sentences preceding (PREC) and following (FOLL) a matched sentence.

[5] http://lucene.apache.org/solr)

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

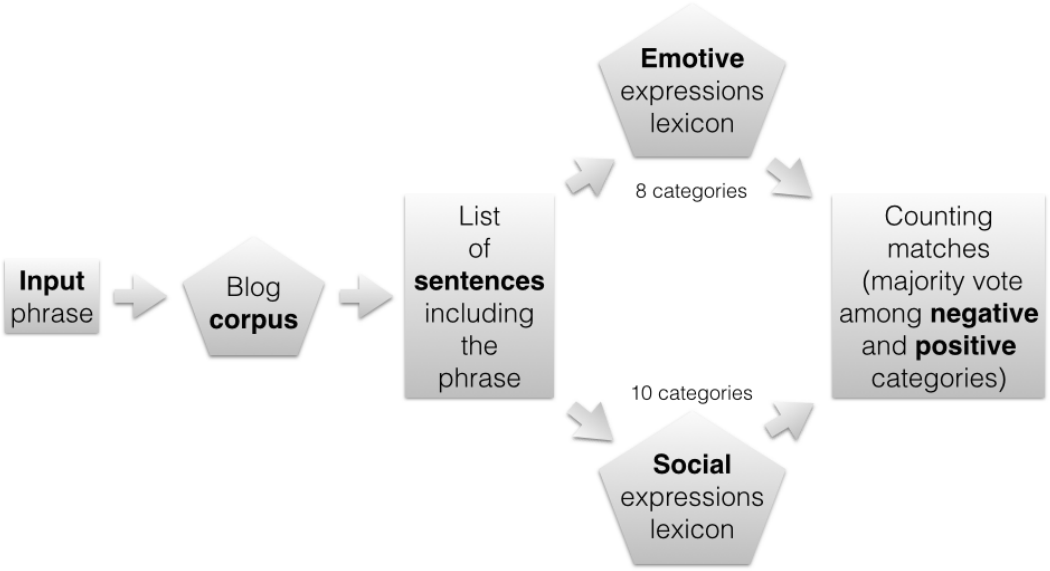

Figure 2: Part of the system that was used to confirm if proposed changes to the baseline system were effective (Instincts Recognition Module was ommitted).

3.1 Human Evaluation

In our previous study [Rzepka and Araki, 2012b] asked 7 Japanese students (22-29 years old, 6 males and one female) to rate 68 input acts on an 11 point morality scale where -5 is the most immoral and +5 is the most moral. 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. But we marked both "no ethical valence" and "context dependent" as ambiguous (AMB) in sake of easier processing, and also because we observed that almost every human act can have moral connotation in a particular context. On 68 evaluations there were only two explicit disagreements between subjects (when evaluating "revenging oneself" and "going to a love hotel") and they 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. These ambiguous acts are problematic because they heavily depend on context and show how different attitude toward a survey a subject can have. Some of them treated the inputs lightly and used common associations (e.g. "driving a car" is a mean for *work* or giving oneself and other people *pleasure*, so should be considered moral), others tended to imagine negative sides of acts (i.e. "driving a car" can surely cause *harm* to people); there were also subjects who always thought about two sides of an act: because there are people in the world who think about "eating pork" as unethical it is safer to mark "eating a pig" as ambiguous. Because such evaluations get scattered through the scale, we decided to treat neighboring agreements as semi-correct, i.e. when most of the subjects evaluated something as bad and the system (the bloggers, to be exact) chose ambiguity, we counted it as 0.5, a value between full agreement (subjects' "bad" evaluated as "bad" and "good" as "good") which gets 1 point, and full disagreement ("explicit error") where the system judged an act as "good" while it was "bad" for most of the subjects (0 points).

3.2 Newly added inputs

To see if a bigger number of inputs (not only new features and search engine) can also change the results, we asked two male information science students (both 23 years old) to create new inputs and to evaluate morality of the other party's set. We newly acquired 61 and 78 inputs and after deleting doublets added them to the previous 68 inputs creating three sets: a) baseline one (68 phrases); b) middle size one (127 phrases) and c) bigger one (207 phrases). The subjects were shown the previously used set as an example, which caused that many of new inputs were variations of existing ones. However we came to the conclusion that it does not make the task easier - in fact, many of the new phrases were rare just because of this tendency to mimic or mix the existing ones. For instance "eating beef" inspired by "eating a cow" could surely be helpful for the system's recall but "eating a car" inspired by "driving a car" and "eating a cow" obviously could not.

3.3 Experimental Results

We ran the system altering data sets and parameters to observe change in the results. The first factor we investigated was the influence of matching lexicon words also in sentences that precede (PREC) and follow (FOLL) the conditional one (which contains the query formed by noun, particle and verb in if-form) to decide if they should be used or not. Small and medium sets showed that using both helps keeping relatively high precision in case of strict evaluation (where neighboring agreement is counted as 0, not 0.5) and semi-correct evaluation (neighboring agreement counted as 0.5) but when all 207 inputs were fed to the system, this tendency was not so obvious anymore and we chose only following sentences to

		68 inputs			127 inputs			207 inputs		
		strict	semi	loose	strict	semi	loose	strict	semi	loose
Minority 20% Majority 80% (No PREC With FOLL)	precision	0.462	0.721	0.981	0.522	0.754	0.986	0.504	0.748	0.991
	recall	0.264	0.359	0.432	0.222	0.292	0.351	0.220	0.295	0.356
	f-score	0.336	0.479	0.600	0.312	0.421	0.517	0.306	0.423	0.524
Minority 30% Majority 70% (No PREC With FOLL)	precision	0.462	0.712	0.962	0.565	0.768	0.971	0.513	0.748	0.983
	recall	0.267	0.359	0.431	0.238	0.298	0.349	0.223	0.296	0.355
	f-score	0.338	0.477	0.595	0.335	0.429	0.513	0.311	0.424	0.522
Minority 33.3% Majority 66.6% (No PREC With FOLL)	precision	0.500	0.731	0.962	0.609	0.790	0.971	0.539	0.761	0.983
	recall	0.283	0.365	0.431	0.251	0.304	0.349	0.232	0.299	0.355
	f-score	0.361	0.487	0.595	0.356	0.439	0.513	0.325	0.429	0.522
Minority 40% Majority 60% (No PREC With FOLL)	precision	0.571	0.765	0.959	0.628	0.797	0.965	0.582	0.768	0.955
	recall	0.298	0.362	0.416	0.303	0.356	0.401	0.241	0.295	0.342
	f-score	0.392	0.492	0.580	0.409	0.492	0.567	0.340	0.426	0.504
Minority 49% Majority 51% (No PREC With FOLL)	precision	0.500	0.712	0.923	0.580	0.768	0.957	0.557	0.748	0.939
	recall	0.289	0.366	0.429	0.244	0.299	0.347	0.242	0.301	0.351
	f-score	0.366	0.484	0.585	0.343	0.431	0.510	0.338	0.429	0.511

Figure 3: Differences in results depending on how the ambiguity margin is set. It appeared that it is better to use 40% vs 60% margin, not the 33.3% and 66.6% one proposed in previous work.

be used in further experiments as it had the highest number of correct and semi-correct moral estimations and the highest loose evaluation which we use mostly to see how many full disagreements a given run produced (it is calculated by giving semi-correct estimations 1 point). The high score for preceding sentences was surprising and gave us ideas for using word lexicons to discover intentions of acts as mentioned earlier in the paper.

The next step was to see if our margin for ambiguity set for the baseline system (below 33.3% is minority and above 66.6% is majority, anything else is ambiguous), we recalculated the agreement between subjects and the system (version using following sentences) in four additional scenarios: a) where minority is below 49% and majority above 51% (no ambiguity allowed); b) where minority is below 40% and majority above 60% (slight ambiguity allowed); c) where minority is below 30% and majority above 70% (big ambiguity allowed) and d) where minority is below 20% and majority above 80% (very big ambiguity allowed). As shown in Figure 3, precision for strict and semi-strict evaluation is best when the ambiguity margin is smaller (40–60) than "1/3 positive 1/3 ambiguous 1/3 negative" approach used in previous research. However, because the differences are not that significant, we need to continue experiments with different ranges, sets of inputs and increase the number of evaluators for the new acts. Naturally enlarging ambiguity margin (see red frames in Figure 3) is beneficial for decreasing explicit errors because it is quite uncommon that the system and subjects diametrically. Why they still occur, we explain later in the error analysis section.

The final comparison we performed was to see if we managed to keep the precision without decreasing f-score and at the same time to decrease number of explicit errors which are not so context dependent as semi-correct ones. We used the same set-up as in [Rzepka and Araki, 2012b] using 68 examples and ambiguity between 33.3% and 66.6%. It appeared that the proposed system with an enhanced Nakamura and Kohlberg-based lexicons had a slightly lower f-score than the original (0.445 vs. 0.467) but decreased the number of full disagreements from 6 to 2 (all the 6 problematic judgments were automatically evaluated as correct or semi-correct, however two completely new "explicit errors" appeared). By enhancing the lexicon we mean deleting some Chinese characters which caused problems as noticed in [Rzepka and Araki, 2012b] and heuristically adding more keywords describing social consequences, although they were not so much inspired by the Kohlbergian theory of moral development. For instance, "being arrested" usually is a consequence of some unsocial behavior, but "being killed" is just a very bad outcome with wider recall than "being executed", and adding it appeared beneficial for the overall performance.

4 Error Analysis

Most of the errors in our experiments were caused by insufficient context processing, which is our current work. Having said so, two explicit errors from the system comparison show that our system did more or less a sufficient job of recognizing bad from good, it just had insufficient number of example sentences retrieved. One error came from specific tendencies in the found blog entries, another from weaknesses of the setup. The first input judged differently by the system and the human subjects was "preventing conception". It ap-

peared that most of the bloggers expressing consequences of this act were writing about how their pets reacted to it. It is not unusual for Japanese to write about their pets on the Internet, and this example perfectly illustrates the urgent need for further semantic analysis that would help recognizing e.g. agents and patients of acts. It can be done by one of the tools we already use – semantic tagger ASA, but it needs further improvements because of still noisy analysis. When it is used, there is also a need to decide what categories of agents (act doers, actors) and patients (act receivers, objects) should be used. Simple human / animal / object categorization is insufficient, because, as Bentham already noticed centuries ago, ones mother's ethical value is definitely different when compared to the value of a complete stranger. The second sentence that had an opposite polarity was "avoiding / preventing a war". The noise was caused by small number of examples and one sentence saying *for preventing wars it is not enough to say our children that 'war is a misery', 'war is a tragedy' and so on!* that had two negative keywords (*misery* and *tragedy*) which tipped the scales in favor of negative consequences. Firstly, the sentences with citations should probably be ignored for now as ones with bracketed explanations. Secondly, the if-form "tame" that was used here, has also meaning of purpose in Japanese language, hence we may need to look closer at the particular forms and see their individual performances.

5 Conclusions and Future Work

In this paper we presented our method of using opinion-mining and sentiment analysis for retrieving written descriptions of bloggers' experiences and opinions in order to equip a machine with knowledge about common consequences of simple acts. This simplicity is making the processing easier, but also causes errors that need to be dealt with by the system in the near future. Here we introduced the first step that moves baseline research from the most shallow keyword matching level into the version that handles negations, conditionality and uses less ambiguous phrases for recognizing polarities of consequences. We discovered that many words taken from Nakamura dictionary of emotive expressions cause noise and the lexicon of words inspired by Kohlberg could be extended with new phrases that can be discovered from the search results. For instance sentences describing lives of women who had to "sell their bodies" bring new descriptions of tragic lives, and these descriptions could be reused as a lexicon entries if they appear often in strongly negative sentences. Also, using "selling a body" input example, the reasons for desperation could tell the machine more about a difference between "doing it to become a Hollywood star" and "doing it to feed her children". This influence of context is crucial for arriving at fair conclusions, and as mentioned earlier, to deepen the analysis we are already working on applying vectors of Bentham's Felicific Calculus [Bentham, 1789] to our system (they deal with, among others, time, scale, continuity of pain and pleasure). In this paper we showed that even a few techniques for more thorough analysis, although more time consuming (4.38 seconds on average for one judgement), can provide precision without Google size indices. Nonetheless, when it comes to implementing our method on a moral dialog agent, low recall is one of the biggest drawbacks of our system. But since it is difficult for academia to utilize super fast computers and millions of gigabytes of disk space, we plan not only to crawl more texts but also to use other existing NLP methods in order to improve the recall. For instance, none of the sentences extracted by "killing a president" input, contained a lexicon keyword. This can be improved by finding more sentences with synonyms of "president" or by using a second layer of search where chunks of sentences describing murdering a country's leader become inputs themselves.

Another important future work is further research on evaluating moral decisions, and tests with different types of surveys. Shortly stated acts, even with a micro-context of what or who is an object, patient, target, etc. are difficult to judge. Every subject, with his or her baggage of experiences, imagines things differently; therefore probably it would be more natural to utilize micro story-type questionnaires that enable subjects to grasp details of the morally ambiguous situation. But for that, we need to expand input and search processes, which will take some time.

Until recently, WWW was treated as a massive garbage can full of sex and violence which is not useful for intelligent machines, especially for AMAs. With this paper we want to catch the cognitive computing community's attention to the fact that computers with constantly improving NLP tools and a tiny involvement from human (258 keywords divided into two categories) are capable of replacing or supplementing physical perception until artificial five senses are able to learn from the real world. Our algorithm was able to filter out meaningless noise and read stories of people whose majority, surprisingly for many, seems to represent healthy common sense, which can further be used for applications of existing or newly created moral solvers and advisors. Furthermore, although packed with descriptions of unreal worlds and games where killing is a pleasant purpose, WWW becomes a "knowledge soup" [Sowa, 2004], a part of Global Brain [Heylighen, 2011] from which machines are slowly learning how to distinct fantasy from more realistic stories and to avoid assuming that people can fly on broomsticks because Harry Potter can. But even if bloggers create different fantasy worlds, they share their human emotions, describe punishments for evil deeds and the empathic brains react to happy and unhappy moments also of aliens and dragons. We see it as a chance for acquiring a useful knowledge on what we care about and on what we would do if the object of our care faced danger.

We conclude again disagreeing with [Kaku, 2014] who claims we do not have machines that can simulate the future, not about our everyday life. We think we are quite close – by guessing possible consequences they could stop us from doing bad things (or at least try to) or report unethical behavior in a near future. Presented system can be easily implemented on existing, simple robotic systems as Roomba vacuum-cleaner or be added to wearable devices as cellphones or glasses. Even without context processing it could extract act chunks of our utterances and consult our ideas with the common sense of a thousand of other people

and warn us if we are possibly victims of an cognitive bias. In the future, when NLP techniques become more advanced, we are planning to add a doubt mechanism for the retrieved common sense by confronting acquired knowledge with results retrieved from blogs in other languages, scientific paper repositories and legal documents. We have already started a project which will try to detect statistical mistakes we usually make an example of various errors described by [Tversky and Kahneman, 1974] in their lifetime work [Kahneman, 2003].

References

[Anderson and Anderson, 2014] Michael Anderson and Susan Leigh Anderson. Geneth: A general ethical dilemma analyzer. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, July 27 -31, 2014, Québec City, Québec, Canada.*, pages 253–261, 2014.

[Bentham, 1789] Jeremy Bentham. *An Introduction to the Principles and Morals of Legislation*. T. Payne, London, 1789.

[Bottou, 1998] Léon Bottou. Online Algorithms and Stochastic Approximations. In *Online Learning and Neural Networks*. Cambridge University Press, 1998.

[Carlson *et al.*, 2010] A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E.R. Hruschka Jr., and T.M. Mitchell. Toward an architecture for never-ending language learning. In *Proceedings of the Conference on Artificial Intelligence (AAAI)*, pages 1306–1313. AAAI Press, 2010.

[Guarini, 2006] Marcello Guarini. Particularism and the classification and reclassification of moral cases. *IEEE Intelligent Systems*, 21(4):22–28, July/August 2006.

[Heylighen, 2011] Francis Heylighen. *Evolution: Cosmic, Biological, and Social*, chapter Conceptions of a Global Brain: an historical review, pages 274 – 289. Uchitel Publishing, 2011.

[Kahneman, 2003] D. Kahneman. A perspective on judgement and choice. *American Psychologist*, 58:697–720, 2003.

[Kaku, 2014] Michio Kaku. *The Future of the Mind: The Scientific Quest to Understand, Enhance, and Empower the Mind*. Doubleday, first edition, February 2014.

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

[Kudo, 2005] Taku. Kudo. Mecab : Yet another part-of-speech and morphological analyzer. *http://taku910.github.io/mecab/*, 2005.

[Lenat and Guha, 1989] Douglas B. Lenat and R. V. Guha. *Building Large Knowledge-Based Systems; Representation and Inference in the Cyc Project*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1st edition, 1989.

[Liu and Singh, 2004] Hugo Liu and Push Singh. Conceptnet: A practical commonsense reasoning toolkit. *BT Technology Journal*, 22:211–226, 2004.

[McDougall, 1923] W. McDougall. *Outline of Psychology*. Charles Scribner's Sons, 1923.

[McLaren, 2003] Bruce M. McLaren. Extensionally defining principles and cases in ethics: An {AI} model. *Artificial Intelligence*, 150(1–2):145 – 181, 2003. {AI} and Law.

[Miller, 1995] George A. Miller. Wordnet: A lexical database for english. *Communications of the ACM*, 38:39–41, 1995.

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

[Ptaszynski *et al.*, 2012] Michal Ptaszynski, Rafal Rzepka, Kenji Araki, and Yoshio Momouchi. Annotating syntactic information on 5 billion word corpus of japanese blogs. In *In Proceedings of The Eighteenth Annual Meeting of The Association for Natural Language Processing (NLP-2012)*, volume 14-16, pages 385–388, 2012.

[Rzepka and Araki, 2012a] Rafal Rzepka and Kenji Araki. Language of emotions for simulating moral imagination. In *Proceedings of The 6th Conference on Language, Discourse, and Cognition (CLDC 2012)*, May 2012.

[Rzepka and Araki, 2012b] Rafal Rzepka and Kenji Araki. Polarization of consequence expressions for an automatic ethical judgment based on moral stages theory. Technical report, IPSJ, 2012.

[Schank and Abelson, 1977] R. Schank and R. Abelson. *Scripts, plans, goals and understanding: An inquiry into human knowledge structures*. Lawrence Erlbaum Associates, Hillsdale, NJ., 1977.

[Settles, 2009] Burr Settles. Active learning literature survey. Technical Report 1648, University of Wisconsin–Madison, 2009.

[Singh *et al.*, 2002] Push Singh, Thomas Lin, Erik T Mueller, Grace Lim, Travell Perkins, and Wan Li Zhu. Open mind common sense: Knowledge acquisition from the general public. In *On the Move to Meaningful Internet Systems 2002: CoopIS, DOA, and ODBASE*, pages 1223–1237. Springer, 2002.

[Sowa, 2004] John F. Sowa. The challenge of knowledge soup. In *In Proc. First International WordNet Conference*, page 15, 2004.

[Suchanek *et al.*, 2007] Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. Yago: A Core of Semantic Knowledge Unifying WordNet and Wikipedia. In *Proceedings of the 16th international conference on World Wide Web*, WWW '07, pages 697–706, New York, NY, USA, 2007. ACM.

[Takeuchi *et al.*, 2010] Koichi Takeuchi, Suguru Tsuchiyama, Masato Moriya, and Yuuki Moriyasu. Construction of argument structure analyzer toward searching same situations and actions. Technical Report 390, IEICE technical report. Natural language understanding and models of communication, jan 2010.

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