Conscious vs. Unaware Evaluation – – Using Collective Intelligence for an Automatic Evaluation of Acts

Rafal Rzepka and Kenji Araki
Hokkaido University, Japan
{rzepka,araki}@ist.hokudai.ac.jp

Abstract

In this paper we introduce our idea how non-restricted text-based knowledge bases can be used for simulating human evaluators. To illustrate this idea we introduce an example in which independent Internet resources help to automatically evaluate a given act by recognizing polarity of its possible outcomes described in various natural language corpora. We present results showing how the source size, the majority-minority proportions and utilized lexicons influenced automatic moral judgement and how it differed when compared with human subjects. With our paper we would like to spark a discussion on alternative, human knowledge-based AI self-evaluation methods which are understandable by humans.

1 Introduction

As Artificial Intelligence is gradually being spread from highly specialized tasks to more general ones [Goertzel and Pennachin, 2007; Hutter, 2007; Arel *et al.*, 2010], its testing and evaluation become problematic. Various approaches were proposed to tackle this problem [Hernández-Orallo, 2016], however, most, if not all of them, remain only theoretical or narrow. Our idea has derived from the experience with non-task oriented dialog systems and (in the current state) is based only on language and requires a verbal (textual) description of the action and its result. Because it is not natural to ask an interlocutor (or the supervisor) for feedback every time a dialog system processes or generates an output, we have been using Internet text data as a simulation valuating level of common sense or naturalness also on the semantic, not only lexical, level [Rzepka *et al.*, 2005; 2010]. We also utilize similar approach in command understanding by robots [Takagi *et al.*, 2011], where just using Twitter logs a robotic vacuum-cleaner guesses not only indirect suggestion as "this room is dirty" but also is able to deduce that it cannot clean bathtubs (as no example of similar usage was found in the corpus). We realize that such a method is imperfect but quite natural when it comes to human "commonsensical" knowledge-based way of intelligence assessment. From the verbal description of somebody's act we can attempt to recognize if it was meaningful, fast, not to overcomplicated, harmful, etc. There are several problems with this approach. One is the insufficient description of an act in case of language. This description can be also biased if subjective adjectives are used instead of e.g. exact measurements. Another natural objection is the imperfectness of the human being, even if so called Wisdom of Crowd phenomena is used. We assume that in future most of machine learning-based AI systems will be able to describe recognized / learned details in natural language helping the users, and other system as ours, to enrich the input. For time being judges of AGIs could be using objective narration as "System A performed task X in three hours" or "System B tried to perform task Y using scissors". The example of moral evaluation given in this paper shows how our method concentrates on (e.g. legal) consequences and people emotional reactions to avoid (at least to some extent) subjectivity while judging an act.

2 Automatic Morality Evaluation Task

It has been widely discussed lately if an autonomous agent can and should perform ethical decisions. In our approach [Rzepka and Araki, 2005], an agent indirectly asks Internet users for their experiences and evaluates a given act by automatic retrieval of descriptions of emotional reactions to this act (e.g. "he hit her" $\rightarrow$ "she cried") and social consequences ("he hit her" $\rightarrow$ "she sued him"). This simple idea is based on an assumption that even if the majority of people behave immorally from time to time, they tend to correctly judge others when the third person misbehaves. In the past we were able to reach almost 80% agreement with human subjects but these results included "semicorrect" judgements where ambiguous evaluations were treated lighter than definitely incorrect ones (judgement opposite to majority of humans). In this research we managed to reach accuracy over 85% even without including scores for ambiguous automatic evaluations thanks to introducing new corpora. As we show later in the paper, not only a size but also type of a given corpus seems to be important for the quality of retrieved knowledge which can be used for automatic evaluation. Our research question is: "if given a sufficient data describing similar situations with their consequences being described, can an autonomous agent judge a situation using the majority vote"? Majority, as we show later, is problematic and there are cases like euthanasia where the difference in polarity of reactions is not distinct enough. Because data other than text is still diffi-

cult to be automatically analyzed due to insufficient acts and consequences recognition, the question is unanswered. Also natural language understanding techniques and textual data, although constantly growing, are often insufficient for more complicated (contextual) input and search. Having stated that, negative consequences not only of "stealing" but the difference in weights between "stealing a car" and "stealing an apple" can be discovered from the textual resources and an agent can decide what kind of utterance or action it should take upon this evaluation of "unaware" Internet users. Below we report how different such evaluation is when compared to "conscious" survey respondents asked to evaluate the same set of acts.

2.1 State of the Art

Our research is a crossing of common sense knowledge acquisition, sentiment analysis and machine ethics, therefore it is rather difficult to report the whole spectrum of related work. Research on AMAs (Autonomous Moral Agents) is mostly theoretical, and except ours, there is only few systems that deal with a wide range of situations. GenEth, the learning system developed by [Anderson and Anderson, 2014], is in theory able to learn from ethicist's decisions how to judge novel inputs. However, the supervising process would be very laborious and costly and indefinite number of contextual conditions could cause problems not only for the supervisors but also for the learning itself. The SIROCCO system [McLaren, 2003] also utilizes case-based reasoning on examples from professional engineering ethic cases in order to help predicting 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. [Guarini, 2006] utilizes a simple recurrent network to trained a system that uses sentences about killing and allowing to die described as acceptable or unacceptable, however his system requires manual labelling of learning data. When it comes to common sense knowledge acquisition [Suchanek *et al.*, 2007; Carlson *et al.*, 2010; Speer and Havasi, 2012] and sentiment analysis [Cambria *et al.*, 2013], we use classic lexicon-based approach to avoid costly and laborious data preparations for machine learning[1].

2.2 Summary of Previous Trials with Ethical Judgement

Details of our previous systems, lexicons and experiments are presented in [Rzepka and Araki, 2012; 2015] but in the section we briefly describe some important details. We have been using seven lexicons (details in the next section) but using only one kind of corpus which is a Japanese[2] blog site snapshot [Ptaszynski *et al.*, 2012]. Our system simply searches for sentences describing an act and retrieves as many as possible. Then it matches all positive and negative words in the right side of the act phrase as it is where consequences are usually described (reasons naturally tend to be before the act). For example, if an analyzed input act was "to hit a girl" (see Table 1 for examples of acts used in this research), our system could find the following sentence in a blog or Twitter:

He hit the girl so hard that she almost died, terrible!

Lexicons used for sentiment analysis usually consist of words and phrases as "die" or "terrible" labeled as negative and just by comparing counts of positives and negatives, the algorithm can evaluate the act. It is discussable if such a utilitarian approach is the best one, but to our best knowledge it is the only attempt to cover such wide range of inputs (basically any act given in a natural language is processable). More details are given in the "System Overview" section.

We have managed to increase accuracy of the automatic moral judgements by adding if-forms to verbs in input, however, we did not perform any experiments with different corpora. All utilized knowledge sources are presented in the next section. As for the lexicons, they are divided into positive and negative phrases and are introduced in a separate subsection.

3 Data Used

3.1 Text Knowledge Resources

To the Ameba blog corpus [Ptaszynski *et al.*, 2012], we have tested five additional corpora. The closest one is "Random WWW" corpus generated using a search engine and most common Japanese words[3]. Instead of limiting knowledge to blog entries, it covers pages without any restrictions but lexical. Similarly, Google N-gram[4], the biggest corpus we used, also provides data indexed by crawlers. However, the corpus is divided into grams (morphological chunks in case of Japanese) and does not contain long sentences. It means that lexicon phrases can be found only on the very limited space as our system searches for these phrases after the input act. Two other sets are collected by the authors. The first is made from Internet Relay Chat (IRC) open channels logs collected from 1999 till 2009 and the second is made from tweets saved in 2010. It must be noted that IRC logs contain channel operating messages, not only natural language messages; non-Japanese utterances were not removed. The last corpus we utilized is Aozora Bunko[5], freely available repository of Japanese literature and poetry which is not limited by copyrights. The texts are annotated with readings and not all annotations were removed, many citations which were not using periods became big text chunks. Still, as all entries in every corpus, we call a line of corpus a *sentence* and we do not set length boundaries as we did in previous experiments to avoid wrongly divided blog entries which used emoticons instead of periods. Sizes and ratios are shown in Table 2.

[1] We plan to implement machine learning as soon as the accuracy of our automatic knowledge retrieval is sufficient to generate valuable datasets automatically

[2] We work on Japanese data and all examples in this paper are translations.

[3] http://corpus.leeds.ac.uk/internet.html

[4] https://research.googleblog.com/2006/08/all-our-n-gram-are-belong-to-you.html

[5] http://darthcrimson.org/digital-japanese-literature-aozora-bunko/

Table 1: Examples of 68 input acts.

accepting a bribe	helping a friend	making one drink alcohol	stealing a bicycle	being unfaithful
avoiding war	hiding a crime	making a minor drink alcohol	stealing a car	killing a bacteria
becoming an egoist	hiring a foreigner	buying a prostitute	taking one's girlfriend	revenging oneself
being deceived	hurting somebody	performing mercy killing	taking one's money	killing many
being fired	ignoring a crime	preventing conception	throwing away bread	having sex

Table 2: Ratios of corpora sentences containing 68 input acts from the experiment.

Corpus name	Sentences (total)	Acts found per ten thousand sentences	Sentences with acts
Internet Relay Chat	4,155,193	0.48	198
Books	7,227,443	1.61	1,164
Random WWW	12,759,191	6.02	7,687
Twitter	79,586,416	0.82	6,538
Blogs	341,400,776	0.97	32,981
Google 7gram	570,204,070	0.49	27,716
CombCorp	1,015,333,089	0.01	76,284

4 System Overview

We prepared a simplified version of our previously described systems because of time and resources constrains. Algorithm, shown in Figure 1, finds input acts in a corpus and after retrieving sentences with these acts, matches consequences on the right side of an act (as reasons are more often on the left side of an act and outcomes come after, later in the sentence). Then a majority[6] decides if the corpus judgement is "Correct" (above majority threshold), "Incorrect" (below minority threshold) or "Ambiguous" (between minority and majority thresholds). The test data for comparison with human evaluators was created seven Japanese students (22-29 years old, 6 males and one female) who rated 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. We marked both "no ethical valence" and "context dependent" as "Ambiguous".

4.1 Lexicons

We have decided to reuse all previously used lexicons for discovering emotional and social consequences of human acts. The first two use emotive expressions collected by [Nakamura, 1993] – one contains all phrases collected from Japanese literature, the second is limited to the most often used phrases from the first (we utilize the shorter version when processing speed is needed, e.g. in dialog systems). The third set is based on the Kohlbergs theory of moral development [Kohlberg, 1981] and was created by the authors manually by choosing related words from WordNet. Phrases like "be scolded" (negative) or "be awarded" (positive) are examples of what we call *social consequences*. These consequences, combined with emotional ones from Nakamura, became EmoSoc corpus, the one which scored highest in previous experiments. Version used for the first experiments [Rzepka and Araki, 2012] is called "EmoSocOld" and the newer version used in [Rzepka and Araki, 2015] became "EmoSoc". The former lacked important phrases from Nakamura's "like" and "dislike" categories, and missing "hate"-related words were added. We also use a corpus generated by machine learning algorithm by [Takamura *et al.*, 2005] meant for opinion mining and sentiment analysis tasks of Japanese language. Their method assigned a real value in the range -1 to +1, where the words assigned with values close to -1 are supposed to be negative, and the words assigned with values close to +1 are supposed to be positive. However, our preliminary experiments with this set showed that the closer the values are to zero, the more noise it causes, so we took only the most distinctly positive and negative keywords, leaving only 5,756 expressions out of 55,125 (ones with value higher than 0.9 and lower than -0.9). Other lexicon we used for comparison is called "JAppraisal"[7] and we used its full set of 9,590 words divided into positive and negative ones according to Appraisal theory, i.e. a linguistic model of evaluative language.

To see if quantity is also important, we combined all corpora into one corpus (called "CombCorp" herafter).

4.2 Conditions

We did not limit lengths of sentences with act phrases, and the system did not process previous and following sentences. No conditional forms of verbs were used, only stemmed verbs to assure as broadest coverages as possible in a short time, which was crucial as searching six corpora instead of one. We performed two experiments – one which treated any single consequence that was found as sufficient for the judgement

[6] 51%,55%,60%,66.6%,70%,75%,80%,85%,90%,95%,99% thresholds were tested; minority is calculated with 100 minus majority.

[7] http://www.gsk.or.jp/catalog_e.html

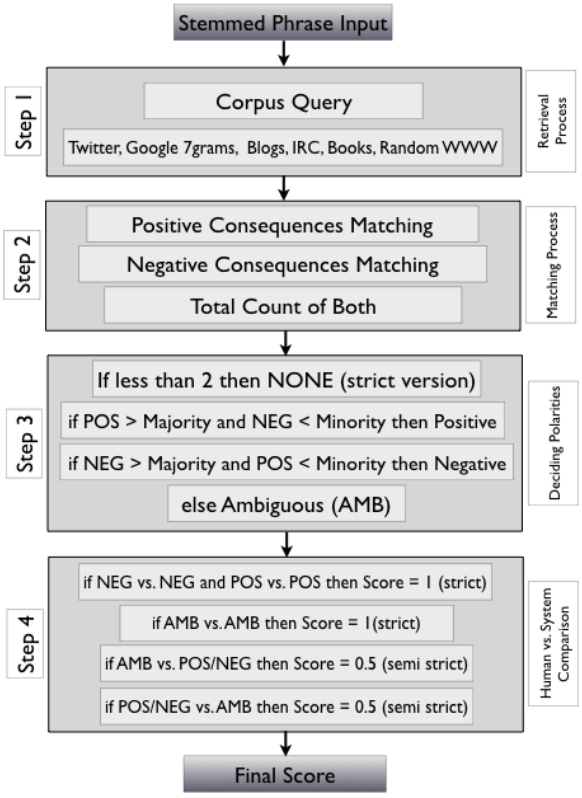

Figure 1: Algorithm for comparing different corpora.

and another where a single case was treated as noise and ignored (meaning that if if there was only one description of a given act, the system treated it as too rare and insufficient).

5 Experiments and Results

We have run 26,928 experiments (68 acts, 6 corpora, 6 lexicons, 11 majority thresholds) twice – first part did not use "at least two consequences" rule, the second utilized "single consequence is sufficient" limitation as previous versions. Without hit restriction, the highest agreement was achieved by EmoSocOld lexicon on 7gram corpus: 77.27%, while adding the simple restriction allowed EmoSoc on 7grams to reach 85.71% accuracy, both with 51-60 majority thresholds (see Table 3 for results after applying the "at least two" rule). As shown in Table 4, the biggest corpus and older version of social and emotional consequences combined scored more than 88% in semi-strict configuration while previous systems never surpassed 80%. Although EmoSoc and Nakamura lexicons scored high, it must be noted that the lowest accuracy was achieved by the Internet Relay Chat corpus also using these two lexicons, especially with high majority thresholds (99 vs. 1 and 90 vs. 10).

Interestingly, the Internet Relay Chat corpus, as shown in Table 1, has almost the same ratio of input act mentioned per ten thousand sentences as Google 7gram corpus (0.48 vs. 0.49 acts per 10.000 sentences with acts). It can mean that Japanese people do not avoid talking about acts but tend to refrain from describing consequences and judge more than when expressing themselves on other media. Accuracy achieved by Books, Blogs and CombCorp corpora gathered in the middle, Twitter and Random WWW closer to the top.

As expected, the biggest corpus (CombCorp) retrieved the biggest number of consequences (57 out of 68) together with one of the largest lexicons – the Takamura dictionary. Also the Blog (Ameba) corpus and the JAppraisal lexicon found most of acts (53-55) but the lexicon size was not assuring higher accuracy as we showed in previous research. The biggest lexicon and combined corpus brought the largest number of correct agreements (35 with 90 majority threshold), but the most incorrect judgements also were brought from CombCorp, especially when accompanied by the JAppraisal (20 with 51 majority threshold). Both the Takamura (16 with 90 majority threshold) and the JAppraisal (also 16 with 66.6 majority threshold) caused the largest number of ambiguous judgements. See Tables 3, 4 and 5 for the best combinations of corpora – lexicon – threshold setup combinations.

Most often corpora were erroneous while judging act of "drinking alcohol" (340 times, many due to the books corpus), although it is discussable if human subjects were correct assigning "good" label to this act not deeply thinking about bad consequences (they were all students, mostly males). Another example showing characteristic tendencies in incorrect judgements is "killing a dolphin" act judged automatically as "good" by 7grams because two sentences (or rather parts of sentences) containing this act and its consequence were too short to discover negations in the end of the full-length originals. Specificity of the stories written online are visible in erroneous judgments of "to cooperate" act. It seemed nice to evaluators without any context, but life is full of bad examples of cooperation. Another problem is shown by other popular misjudgment, "to steal a girlfriend". It seems that many Japanese bloggers are happy about stealing somebody's lover but there are not enough mentions about sad sides of having a lover stolen by somebody else.

6 Discussion

The system presented above is simplistic and limitations of this publication does not allow to describe existing problems and solutions to these problems. It is still discussable which types of intelligence could be evaluated via language but we believe this approach can be at least temporarily used for self-testing in systems which interact verbally with a user. Currently we are testing another web corpus, subtitles corpus, synonyms and we are experimenting with more acts introduced in [Rzepka and Araki, 2015] (although more human evaluators are needed) and [Rzepka *et al.*, 2016] (act worth praising and condemning). We have already started adding more sophisticated context analysis to see how reasons for a given behavior can change the judgement (stealing an apple to feed somebody's little brother vs. just stealing an apple). These additions might be directly useful for evaluating other tasks requiring different types of intelligence. The same can be said about our other projects – recognizing time periods[Rzepka and Araki, 2017], distinguishing physi-

Table 3: Top accuracy scores including ambiguous scoring (0 points for non-absolute errors).

Corpus	Lexicon	Majority vs. Minority	Accuracy
Random WWW	EmoSoc	55 vs. 45	79.16%
Random WWW	EmoSoc	51 vs. 49	79.16%
Google 7grams	EmoSoc	55 vs. 45	84.21%
Google 7grams	EmoSoc	66.6 vs. 33.4	84.21%
Google 7grams	EmoSoc	60 vs. 40	84.21%
Google 7grams	EmoSoc	51 vs. 49	84.21%
Google 7grams	EmoSoc	65 vs. 35	84.21%
Google 7grams	Nakamura	55 vs. 45	84.61%
Google 7grams	Nakamura	51 vs. 49	84.61%
Google 7grams	Nakamura	60 vs. 40	84.61%
Google 7grams	EmoSocOld	51 vs. 49	85.71%
Google 7grams	EmoSocOld	55 vs. 45	85.71%
Google 7grams	EmoSocOld	60 vs. 40	85.71%

Table 4: Top accuracy scores including ambiguous scoring (0.5 points for non-absolute errors).

Corpus	Lexicon	Majority vs. Minority	Accuracy
RandomWWW	EmoSocOld	51 vs.49	88.09%
RandomWWW	EmoSocOld	55 vs. 45	88.09%
Google7grams	EmoSoc	55 vs. 45	89.47%
Google7grams	EmoSoc	66.6 vs. 33.4	89.47%
Google7grams	EmoSoc	60 vs. 40	89.47%
Google7grams	EmoSoc	51 vs. 49	89.47%
Google7grams	EmoSoc	65 vs. 35	89.47%
Google7grams	Nakamura	55 vs. 45	92.30%
Google7grams	Nakamura	51 vs. 49	92.30%
Google7grams	Nakamura	60 vs. 40	92.30%
Google7grams	EmoSocOld	51 vs. 49	92.85%
Google7grams	EmoSocOld	55 vs. 45	92.85%
Google7grams	EmoSocOld	60 vs. 40	92.85%

Table 5: Setup combinations for the best accuracy for all corpora.

Corpus	Lexicon	Majority vs. Minority	Accuracy
Internet Relay Chat	Jappraisal	51 vs. 49	35.0%
Books	EmoSocOld	75 vs. 25	66.66%
Twitter	EmoSoc	70 vs. 30	68.96%
Blogs	EmoSocOld	75 vs. 25	69.44%
CombCorp	EmoSocOld	70 vs. 30	70.45%
RandomWWW	EmoSoc	51 vs. 49	79.16%
Google7grams	EmoSocOld	60 vs. 40	85.71%

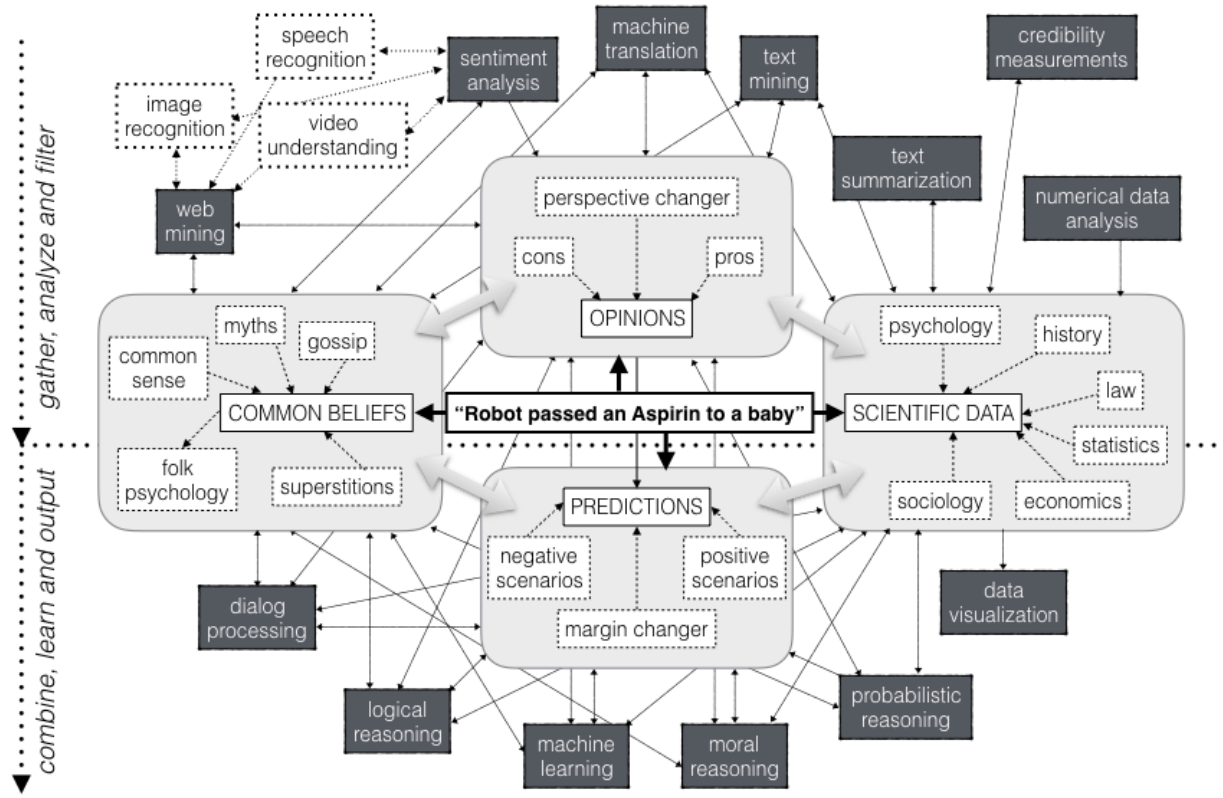

Figure 2: Overview of tasks necessary for extending knowledge and fuller recognition of a given act.

cal and non-fictional objects from abstract and fictional ones, building a script-like knowledge chains database (overview of the tasks being tackled by our group is shown in Figure 2). In our opinion research on knowledge acquisition (both general and specialistic) necessary for testing and evaluating AGIs should not lag behind the algorithms being proposed for this task and this is our message to the researchers working on various aspects of AGI evaluation.

7 Conclusions

In this paper we presented our idea on how automatic evaluation of acts for general purpose agents could work like without any sophisticated algorithm but with vast (still very unordered, uncontrolled and shallow) knowledge. Before any working alternatives are proposed, we utilize natural language processing for automatic knowledge retrieval which is often helpful for simulating human-like estimation of intelligence in our dialog systems. Instead of programming complicated rules, utilizing machine learning, etc. we keep investigating how efficient is borrowing human experiences (as gathering its own is costly and time-consuming for physical robots) for achieving "common sense" (or "common knowledge")-based evaluation of acts (performed either by humans or machines). To illustrate this approach we described our latest tests with the simplest possible matching algorithm using various textual resources and by searching positive and negative consequences of ethically problematic (and non-problematic) acts. We have confirmed that in case of lexicons, quantity is not more important than quality (cleaner Web corpora scored higher), however the size of used knowledge base (in our case text corpora) does matter. Nevertheless, lower accuracy scores of corpus combined from all other corpora suggests that we also need to be careful with choosing a type of corpus and automatically measure credibility of sources. As described in the Discussion section, fuller verbal evaluation of intelligence requires tackling many obstacles and we believe that similarly to the necessity of combining different algorithmic approaches to general purpose AI, the same can be said about its evaluation methods. However, the knowledge required by these algorithm should not be neglected.

8 Acknowledgments

This work was supported by JSPS KAKENHI Grant Number 17K00295. We thank the Centre of Excellence for the Dynamics of Language at Australian National University for providing an ideal environment for multidisciplinary discussion while writing this paper.

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.

[Arel *et al.*, 2010] Itamar Arel, Derek C Rose, and Thomas P Karnowski. Deep machine learning-a new frontier in artificial intelligence research [research frontier]. *IEEE computational intelligence magazine*, 5(4):13–18, 2010.

[Cambria *et al.*, 2013] Erik Cambria, Bjorn Schuller, Yunqing Xia, and Catherine Havasi. New avenues in opinion mining and sentiment analysis. *IEEE Intelligent Systems*, 2(2):15–21, 2013.

[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.

[Goertzel and Pennachin, 2007] Ben Goertzel and Cassio Pennachin. *Artificial General Intelligence*, volume 2. Springer, 2007.

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

[Hernández-Orallo, 2016] José Hernández-Orallo. Evaluation in artificial intelligence: from task-oriented to ability-oriented measurement. *Artificial Intelligence Review*, Aug 2016.

[Hutter, 2007] Marcus Hutter. Universal algorithmic intelligence: A mathematical top→down approach. In B. Goertzel and C. Pennachin, editors, *Artificial General Intelligence*, Cognitive Technologies, pages 227–290. Springer, Berlin, 2007.

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

[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.

[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, 2005] Rafal Rzepka and Kenji Araki. What statistics could do for ethics? - The idea of common sense processing based safety valve. *AAAI Fall Symposium on Machine Ethics, Technical Report FS-05-06*, pages 85–87, 2005.

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

[Rzepka and Araki, 2015] Rafal Rzepka and Kenji Araki. *Rethinking Machine Ethics in the Age of Ubiquitous Technology*, chapter Semantic Analysis of Bloggers Experiences as a Knowledge Source of Average Human Morality, pages 73–95. Hershey: IGI Global, 2015.

[Rzepka and Araki, 2017] Rafal Rzepka and Kenji Araki. Natural language processing for predicting everyday behavior with and without time and duration information. In *Proceedings of the International Symposium on Forecasting IFS 2017, June 25-28, Cairns, Australia.*, 2017.

[Rzepka *et al.*, 2005] Rafal Rzepka, Yali Ge, and Kenji Araki. Naturalness of an utterance based on the automatically retrieved commonsense. In *Proceedings of IJCAI 2005 - Nineteenth International Joint Conference on Artificial Intelligence, Edinburgh, Scotland*, pages 996–998, August 2005.

[Rzepka *et al.*, 2010] Rafal Rzepka, Shinsuke Higuchi, Michal Ptaszynski, Pawel Dybala, and Kenji Araki. When your users are not serious - using web-based associations affect and humor for generating appropriate utterances for inappropriate input. *Journal of the Japanese Society of Artificial Intelligence*, 25(1), 2010.

[Rzepka *et al.*, 2016] Rafal Rzepka, Kohei Matsumoto, and Kenji Araki. Praiseworthy act recognition using web-based knowledge and semantic categories. In *25th International Joint Conference on Artificial Intelligence*, page 41, 2016.

[Speer and Havasi, 2012] Robert Speer and Catherine Havasi. Representing general relational knowledge in conceptnet 5. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may 2012. European Language Resources Association (ELRA).

[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.

[Takagi *et al.*, 2011] Keiichi Takagi, Rafal Rzepka, and Kenji Araki. Just keep tweeting, dear: Web-mining methods for helping a social robot understand user needs. In *Proceedings of AAAI Spring Symposium "Help Me Help You: Bridging the Gaps in Human-Agent Collaboration" (SS05)*, 2011.

[Takamura *et al.*, 2005] Hiroya Takamura, Takashi Inui, and Manabu Okumura. Extracting semantic orientations of words using spin model. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 133–140. Association for Computational Linguistics, 2005.