Limiting Context by Using the Web to Minimize Conceptual Jump Size

Rafal Rzepka, Koichi Muramoto, and Kenji Araki

Graduate School of Information Science and Technology, Hokkaido University, S Kita-ku, Kita 14, Nishi 8, Sapporo, Japan
{kabura,koin,araki}@media.eng.hokudai.ac.jp
http://arakilab.media.hokudai.ac.jp

Abstract. In this paper we introduce our ideas on how experiences from real situations could be processed to decrease what Solomonoff called "Conceptual Jump Size". We introduce applications based on commonsense knowledge showing that vast corpora are able to automatically confirm the validity of the output, and also replace a "trainer", which could lead to decreasing human influence and speeding up the process of finding solutions not provided by such a "trainer" or by real world descriptions. Following this idea, we also suggest a shift toward combining natural languages with programming languages to smoothen transitions between layers of Solomonoff's "Concept net" leading from primitive concepts to a problem solution.

Keywords: Conceptual Jump Size, artificial trainers, Wisdom of (Web) Crowd, Natural Language Processing.

1 Introduction

In his work on Algorithmic Probability (ALP), Solomonoff often underlined that his approach, strongly influenced by the works of Turing, was to build algorithms that are more universal and independent from human influence[1][2], differing from the approaches as of Lenat[3] or Newell[4]. We share his belief that acquiring concepts of learning on different levels is a shortcut to commonsense reasoning, which constitutes a base for more complicated, high level problem solutions and realizing Artificial General Intelligence (AGI). However, we chose a more real-world data-driven approach.

1.1 Common Sense Knowledge as a Contextual Filter

From the beginning of A.I. history, we have been told that people have commonsense while computers do not. From early childhood, human beings acquire various types of knowledge: about the physical world, social rules, and abstract concepts. When it comes to using these experiences, although being bombarded with large amounts of information while perceiving the world around us, we are able to shadow out the irrelevant data and focus on the situation we face.

D.L. Dowe (Ed.): Solomonoff Festschrift, LNAI 7070, pp. 318–326, 2013.
© Springer-Verlag Berlin Heidelberg 2013

Today we know that Broca's area, a part of our brain responsible for language understanding, also plays an important role in ignoring irrelevant input[5]. We can notice the importance of this context fixation when evaluating commonsense knowledge. The human judges opinions vary depending on how rich their imagination or experiences are. However, in real life situations without much time for elaborate thinking, context awareness limits possibilities to the required minimum. When you see Laika sitting inside Sputnik, your association that *a dog can be used to defend your house* becomes shadowed out and the *dogs can be used for experiments* set becomes stronger, while *cats can be used for catching mice* is kept "switched off". Our minds seem to prioritize related domains and avoid irrelevant areas of knowledge. For this reason, after several unsuccessful attempts to use commonsense knowledge effectively and evaluate it fairly, we decided to use contextual restraints for retrieving concepts by limiting them to situations. We chose "house" (rooms, kitchen, bathroom, etc.) as an experimental environment, furniture and utensils as objects, and family members (plus a robot) as actors. We then performed an experiment for automatic discovery of common and uncommon behaviors. It appears that limiting context can easily prevent oversized Concept Jumps[1] (as explained in subsection 1.3) by decreasing the number of strings to be searched. In this paper, we briefly introduce our trials, showing that vast linguistic resources can be used for *training* as Solomonoff predicted[1]. We also take a step further, suggesting that natural language might be the key to faster concept creation and a faster learning process.

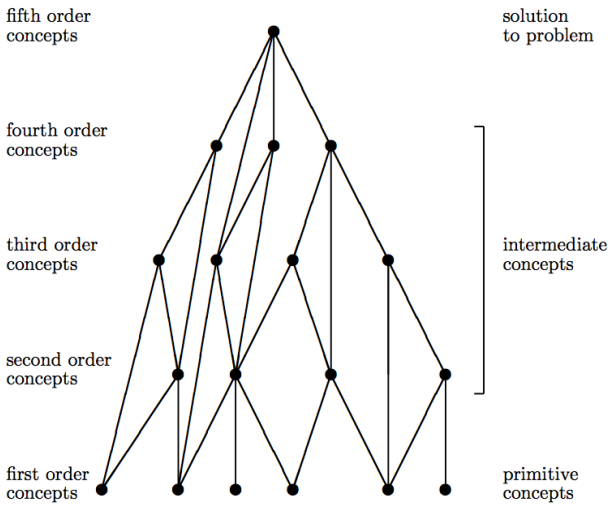

Fig. 1. Simple Concept net introduced by Solomonoff. Our idea is to combine MIT ConceptNet and our Web-based Commonsense Knowledge to smoothen the net processing - its automatic generation and concept manipulation. Preferably by creating some new paradigm which combines natural language based concepts and object-oriented programming language (see Section 3).

1.2 Subjectivity

Solomonoff did not agree with an opinion that subjectivity is "evil" and should not appear in science[2]. He treated finite sample size and model selection error as important sources of error in statistics. In ALP, subjectivity occurs in the latter, which is the choice of "reference" – a universal computer or universal computer language. In the end of his life, Solomonoff was working on an intelligent machine mechanisms that are also visible also in a human infant – born with certain capabilities that assume certain a priori characteristics of its environmenttobe.

> *It expects to breathe air, its immune system is designed for certain kinds of challenges, it is usually able to learn to walk and converse in whatever human language it finds in its early environment. As it matures, its a priori information is modified and augmented by its experience.*

Each time such a system solves a problem or fails, it updates the part of its a priori information that is relevant to problem solving techniques. It resembles a maturing human being and its a priori information grows as the life experience of the system grows. Solomonoff underlines that the subjectivity of algorithmic probability is a necessary feature that enables an intelligent system to incorporate experience of the past into techniques for solving problems of the future. Also in our experiments, especially on language[6] and knowledge acquisition[7] using inductive learning [8], we are driven by similar observations and we often alternate models to see how system's learning performance changes. Output clearly depends not only on the data (environment) it is exposed to but also the choices a researcher makes. Different algorithms, programming languages, users and their behavior always produce different output and this variety of output can lead to the best models but the main problem is that they must be found automatically. To achieve this goal we decided to use large real-world data and mimic what Damasio calls "primordial feelings"[9], an emotion that is a system regulator and the base for rational choices of an universal algorithm. One of he biggest challenges in this stage of our project is the task of limiting context to achieve the shortest possible knowledge testing time.

1.3 What is Conceptual Jump Size

In [1], Solomonoff describes how he uses Levin's search algorithm[10] to deal with the "exponential explosion" problem. Assuming that p_i is the probability of success of the i^{th} trial string of concepts and t_i is the time needed to generate and test that trial, by using the t_i/p_i ordering it is found that for one approximate p_i it is impossible to know t_i before the i^{th} trial, so we cannot make trials in the exact t_i/p_i order. However, it is possible to obtain the t_i/p_i order by selecting a small time limit T, and testing all strings, spending at most a length of time p_iT on the i^{th} string. When a solution is found, the algorithm stops. If not, T is doubled and exhaustive testing is repeated. The process of doubling and testing continues until a solution is found. Solomonoff writes: "It is easy to estimate the total search time needed to discover a particular known solution to a problem.

If p_j is the probability assigned to a particular program, A_j , that solves a problem and it takes time t_j to generate and test that program, then this entire search procedure will take a time less than $2t_j/p_j$ to discover A_j. We call t_j/p_j the "Conceptual Jump Size" (CJS) of A_j"[1]. By using this method we can discover if a machine is practically able to find a particular solution to a problem at a particular state of its development. It is said that CJS is a critical parameter in the design of training sequences and in the overall operation of a system. We hypothesized that using context filtering and Web-based "semantic self-check" could minimize searching time by prioritizing the most obvious clues when a solution is to be found immediately. We introduce some of our experiments, suggesting that this could be a useful shortcut.

2 Our Trials with Commonsense Knowledge

In our research we define "commonsense" quite broadly, by including not only common knowledge of the physical or social world, but also shared beliefs on history, geography or culture. Therefore we allow our programs to retrieve knowledge on famous people or popular events, which is useful especially for dialog systems, when e.g.,a task-oriented mode is suddenly changed by the user's behavior[11]. In the following subsections, we show how such a system can eliminate its semantically erroneous utterances and then how similar ideas can improve the system's handling and generation of concepts.

Self-correcting Universal Dialog System. Non-task oriented dialog systems, usually called chatterbots or chatbots, can be used as free conversational partners that allow the gathering of linguistic information or knowledge about a user for further machine learning, or as a means for dealing with users who have lost their interest in a task of, for example, an automatic information kiosk. The first such conversational system we developed was Modalin, described in detail in [12]. Modalin is a free-topic keyword-based conversational system for Japanese that automatically extracts sets of words related to a conversation topic from Web resources, which was proved to outperform classic ELIZA-like[13] dialog systems and became a successful base for chatbots using emotions[11], humor[14] or causal knowledge[15]. The basic idea of Modalin is simple – after the search engine results extraction process, it generates an utterance, adds modality, and verifies the semantic reliability of the generated phrase before uttering it. Over 80% of the extracted word associations were evaluated as being correct, which was mostly due to an automatic self-correcting process. When a proposition including adjectives, nouns and verbs was created, the system searched for a newly created string on the Internet. If there were only a few such combinations, it discards the candidate and generates a new string with different words. With current search engines operating within seconds it, becomes much easier to avoid "exponential explosion" of meaningless word combinations. However, to search for concepts needed for finding possible solutions, a simple keyword search is not sufficient.

Toward Concept Search and Manipulation. For trials with automatic Shankian-like script retrievals or dialog agents like [16] and [12], the context and usualness are not particularly important, but when it comes to Ambient Intelligence[17] or Machine Ethics[7], the context and unambiguity of results become crucial. For instance, the act of *killing a person* is perceived differently depending on factors like how emotionally close the victim was to the observer, if it happened during a war, or who the victim was to a given society. Solomonoff's"strings" become numerous, long and complicated as they include more elaborate explanations of specific situations. Although the Web is the biggest text resource that exists, there are many problems with retrieving clean and credible knowledge, as many sites use colloquial language which causes noise and makes frequency weights[1] improper. Therefore, we decided to experiment with ConceptNet[18] using the Japanese OMCS[19] database, which is based not on WWW raw data, but on manual input from volunteers. We performed two small experiments to check if a) existing concepts can produce richer and less ambiguous new ones; b) limiting context can help eliminate errors and improve the efficiency of automatic naturalness evaluation of automatically generated concepts.

Generating Chains of Concepts. We tried to generate chains of concepts as follows. First, a random noun is input to ConceptNet, retrieving related concepts. For example, if the input is "a cook", we retrieve *AtLocation*(cook, restaurant), which means that you can nd a cook at a restaurant. Next, "restaurant" is sent back to ConceptNet, and we acquire *AtLocation*(restaurant, department store), as one usually finds restaurants inside department stores in Japan. Finally, we can use this knowledge to create a statement saying one can find a cook inside a department store, or more specifically "at a restaurant inside a department store". We call these combined concepts $Chains(x)$, where x is a number of inputs to ConceptNet. We soon realized that $x = 2$ is probably the maximum which can be useful for joining order levels in Solomonoff's Concept net (see Fig. 1) but often creates nonsense as assumed earlier. To show the scale of the inadequate generation problem, the authors performed a simple evaluation experiment.

Evaluating Concept Triplets. We retrieved pairs of Relations and Concepts using random nouns from the OMCS database for Japanese. From this set we randomly chose one hundred related concepts and evaluated them with a simple scale: "natural", "uncommon" and "unnatural". Only 49% of entries were agreed to be natural relations, 22% were uncommon and 29% unnatural. After an analysis of the data and discussion between evaluators, we agreed that there are at least eight reasons why people label a triplet as "uncommon" or "unnatural".

- Evaluators are not sure about mutual relationship: *AtLocation*(tanker, sea) + *AtLocation*(sea, Yamashita Park). This park is a famous place by the sea,

[1] These weights can be a base for calculating probabilities needed by ALP.

but couples use it for romantic dates and it is not likely you will see heavy ships passing nearby.
- Something is not impossible, but difficult to be evaluated as "natural" by all evaluators. For example $InstanceOf$(salmon, sh) + $AtLocation$(sh, on ship): if the input is "salmon", one rather expects "the sea", or "a plate" or "a fridge" as a natural location for this kind of fish.
- Concepts are obviously related, but the relationship is weak. $PartOf$ (accelerator, car) + $HasProperty$(car, runs on gasoline) would probably score higher if the dependency between using the accelerator and consuming the gasoline was mentioned.
- The OMCS data input by volunteers are not always correct. $PartOf$ (Kunashiri, Japan) + $HasProperty$(Japan, crowded) suggests that there is no conflict regarding whether the disputed island belongs to Russia or Japan. Kunashiri Island also cannot be considered as crowded, as there are very few inhabitants. Only 22% of random triplets were evaluated as "natural", because the more specific the concepts are, the higher the possibility of exceptions.
- Evidently wrong mutual relationship:
$CapableOf$(goose, swim) + $HasSubevent$(swim, wearing swimsuit) suggests that geese swim in clothes.

The analysis showed that 54% of the patterns that scored low were related to context. Therefore we decided to test our ideas about context by narrowing the semantic environment and increasing its density.

Limiting Context. As mentioned in the Introduction, when creating a set of context descriptors we chose a "house" as we are interested in housekeeping robots and such places are often used for commonsense grounding research. We assumed that a machine could name all significant places (we picked up 9 nouns), items (37 nouns) and actions (21 verbs). We designed an algorithm for creating random acts from places, objects and actions. Its basic output was "ACTION with ITEM at PLACE", and this set was sent to Yahoo Japan Blog Search Engine, together with shorter queries, "ACTION in a PLACE" and "ACTION with a TOOL", in order to find frequencies of particular n-grams. The differences between them were used to find the most uncommon semantic components of an action (e.g., eating ice-creams is natural but uncommon if eaten in a bathroom). For this experiment we created the set of context descriptors by hand, but we are currently working on automatic generation of such sets by web-mining techniques supported by knowledge stored in WordNet[20] and ConceptNet. Context is not being labeled (e.g. as house, shop, conference or street); rather, it is based on the top ten semantically significant keywords (actors, place nouns, also description adjectives) and actions that are strongly associated with these keywords. So if an utterer inputs e.g. "That was the best tennis tournament I've ever seen", the context-limiting module retrieves sets of actors (players, spectators, referees, etc.), physical items (balls, rockets, seats, etc.), descriptions (fast, amazing, high-level, etc.) and actions (to play, to watch, to win, etc.) using words from the utterance as queries.

Experiment and Its Results. First, the system itself evaluated permutations of places, items and actions and randomly chosen 100 natural and 100 unnatural self-evaluation outputs, which were shown to three graduate school student evaluators. The self-scoring was done by comparing web frequencies of exact matches. It appeared that the three human evaluators agreed with the system's judgment in 77.08% of cases, showing that context limitation significantly increases accuracy in usualness evaluation.

3 Object-Oriented Programming between Artificial and Natural Languages

The more we work with concepts, the more we realize that they may become instances for joining two realms of language - that is, combining programming and natural languages. If objects, functions, modules or classes were phrases, sentences, instructional stories, etc., the borderline between both worlds could become blurred and Solomonoff's suggestions about training would become easier to realize. As he mentions in [1]:

> *Perhaps the most important kind of training sequence is one that teaches the system to understand English text. By "understand" we mean able to correctly answer questions (in English) about the text. This understanding need not be at all complete, but should be good enough so that ordinary English texts can be a useful source of training for the system. This "training" sequence will involve formal languages of increasing complexity. The first examples of English text will cover a field that the system will already be familiar with – so that it will only have to learn the relationship of the syntax to facts it already knows.*

In our opinion, smoother translation between written natural language and computer-readable logical structures should be faster achieved thanks to enormously growing data[2], which, though seen as very noisy, can help to eliminate a large part of the noise due to its coverage. Engineers from the field of Natural Language Processing (NLP) are making more and more progress in dealing with vast text resources and automatic understanding of these. There are tools (many being improved every month) that help to deal not only with morphological or dependency analysis (at the lexical level) but also with synonyms, homonyms, exceptions, emotive load, usualness and other tasks of the semantic realm. We think that with help of the algorithm community, language engineers could become very helpful for realizing Artificial General Intelligence. NLP is often associated with machine translation, summarization or question answering, which are not associated with simulating mental processes, but the techniques used by these tasks can be easily extended and used for enhancing concept learning and training as Solomonoff proposed.

[2] Nowadays it its mostly text, but one can imagine objects containing videos, sounds or smells.

4 Conclusions

We have described the idea of Conceptual Jump Size as introduced by Solomonoff, and suggested how it can be minimized by vast raw corpora such as World Wide Web textual resources. We introduced algorithms (dialog system and concept generators) where the training process is made by the WWW instead of humans, and showed improvements in their accuracy after using a Web-based self-correcting process. Although there are other research projects on commonsense knowledge enrichment, and also for specified context[21], the main difference is that we aim at universality; i.e., the same system must be able to limit any context in any situation with any kind of agents, objects or places. Finally, we briefly suggested that object-oriented programming and concepts could become a key for creating an intermediate instance between natural and programming languages. We have not introduced any particular algorithm for minimizing Conceptual Jump Size yet, but we believe that our thoughts and experiences on how the lack of settled context makes it difficult to work with common sense knowledge, and how to deal with this problem, may give some clues that may help researchers unfamiliar with NLP to get familiar with a different approach to achieving this goal. With this short introduction of our ideas and latest NLP capabilities, we want to encourage formal language-oriented specialists to cooperate with web-mining and language engineers in the way that agricultural machinery designers work together with soil specialists who know not only different kinds of soil, but also know how to prepare loam. We believe such co-research tendencies could help make Solomonoff's dream of realizing universally intelligent machines[2] become reality.

References

1. Solomonoff, R.: A system for incremental learning based on algorithmic probability. In: Proceedings of the Sixth Israeli Conference on Artificial Intelligence, Computer Vision and Pattern Recognition, Tel Aviv, Israel, pp. 515–527 (1989)
2. Solomonoff, R.: Algorithmic Probability – Its Discovery – Its Properties and Application to Strong AI. In: Zenil, H. (ed.) Randomness Through Computation: Some Answers, More Questions, ch. 11, pp. 149–157. World Scientific Publishing Company (2011)
3. Lenat, D.: Theory Formation by Heuristic Search – The Nature of Heuristics II: Background and Examples. Artificial Intelligence 21(1-2), 31–59 (1983)
4. Newell, A., Simon, H.: GPS, a program that simulates human thought. In: Feigenbaum, E., Feldman, J. (eds.) Computers and Thought, pp. 279–293. McGraw–Hill, New York (1963)
5. Haxby, J.V., Horwitz, B., Ungerleider, L.G., Maisog, J., Ma., P.P., Grady, C.L.: The functional organisation of human extrastriate cortex: a PET-rCBF study of selective attention to faces and locations. J. Neurosci. IR, 6336–6353 (1994)
6. Hasegawa, D., Rzepka, R., Araki, K.: Connectives Acquisition in a Humanoid Robot Based on an Inductive Learning Language Acquisition Model. Humanoid Robots, I-Tech Education and Publishing, Vienna (2009), `http://www.intechopen.com/download/pdf/pdfs_id/6235`

7. Rzepka, R., Komuda, R., Araki, K.: Bacteria Lingualis In The Knowledge Soup – A Webcrawler With Affect Recognizer For Acquiring Artificial Empathy. In: Proceedings of The AAAI 2009 Fall Symposium on Biologically Inspired Cognitive Architectures (BICA 2009), Washington, D.C., USA, p. 123 (2009)
8. Araki, K., Tochinai, K.: Effectiveness of natural language processing method using inductive learning. In: Proceedings of IASTED International Conference Artificial Intelligence and Soft Computing, Mexico, pp. 295–300 (2001)
9. Damasio, A.: Self Comes to Mind: Constructing the Conscious Brain. Pantheon (2010)
10. Levin, L.A.: Universal Search Problems. Problemy Peredaci Informacii 9, 115–116 (1973); Translated in Problems of Information Transmission 9, 265–266.
11. Rzepka, R., Higuchi, S., Ptaszynski, M., Dybala, P., Araki, K.: When Your Users Are Not Serious – Using Web-based Associations, Affect and Humor for Generating Appropriate Utterances for Inappropriate Input. Transactions of the Japanese Society for AI 25(1), 114–121 (2010)
12. Higuchi, S., Rzepka, R., Araki, K.: A Casual Conversation System Using Modality and Word Associations Retrieved from the Web. In: Proceedings of The 2008 Conference on Empirical Methods on Natural Language Processing (EMNLP 2008), Honolulu, USA, pp. 382–390 (2008)
13. Weizenbaum, J.: ELIZA – A computer program for the study of natural language communication between man and machine. Commun. ACM 9(1), 36–45 (1966)
14. Dybala, P., Ptaszynski, M., Rzepka, R., Araki, K.: Activating Humans with Humor – A Dialogue System that Users Want to Interact With. IEICE Transactions on Information and Systems Journal, Special Issue on Natural Language Processing and its Applications E92-D(12), 2394–2401 (2009)
15. Fujita, M., Rzepka, R., Araki, K.: Evaluation of Utterances Based on Causal Knowledge Retrieved from Blogs. In: Proceedings of the International Conference Artificial Intelligence and Soft Computing (ASC 2011), pp. 294–299 (2011)
16. Rzepka, R., Ge, Y., Araki, K.: Naturalness of an Utterance Based on the Automatically Retrieved Common Sense. In: Proceedings of IJCAI 2005 – Nineteenth International Joint Conference on Artificial Intelligence, Edinburgh, Scotland (2005), http://www.ijcai.org/papers/post-0490.pdf
17. Rzepka, R., Araki, K.: What About Tests In Smart Environments? On Possible Problems With Common Sense In Ambient Intelligence. In: Proceedings of 2nd Workshop on Artificial Intelligence Techniques for Ambient Intelligence (IJCAI 2007), Hyderabad, India, pp. 92–96 (2007)
18. Havasi, C., Speer, R., Alonso, J.: ConceptNet 3: a Flexible, Multilingual Semantic Network for Common Sense Knowledge. In: Proceedings of Recent Advances in Natural Languages Processing, pp. 277–293 (2007)
19. Singh, P.: Open Mind Common Sense: Knowledge Acquisition from the General Public. In: Meersman, R., Tari, Z. (eds.) CoopIS/DOA/ODBASE 2002. LNCS, vol. 2519, pp. 1223–1237. Springer, Heidelberg (2002)
20. Fellbaum, C.: WordNet: An Electronic Lexical Database. MIT Press, Cambridge (1998)
21. Gupta, R., Kochenderfer, M.K.: Commonsense data acquisition for indoor mobile robots. In: McGuinness, D.L., Ferguson, G. (eds.) AAAI, pp. 605–610. AAAI Press / The MIT Press (2004)