3E2-07

Automatic General Personality Generation Based on WWW

Rafal Rzepka*1 Kenji Araki*1

*1 Graduate School of Information Science and Technology, Hokkaido University

Seizing the opportunity of JSAI "Near Future Challenge" special session, the authors will try to spark a discussion on the need of combining concepts, methods and approaches into one project which could move the Artificial Intelligence field into the next step of its evolution. In authors' opinion the point that should be the scope of such project is the problem of commonsense processing. The possibilities of automatic retrieval of commonsense and affective reasoning influencing the humans way of commonsensical thinking will be introduced with a metaphorical example of an average personality which hypothetically could be created automatically using web-mining techniques.

1. Introduction

It is obvious for everyone that machines would be regarded as much more clever if they could process information while having the background of commonsense - the knowledge which humans are gaining especially through the first decade of their lives. The amount of data that must be inputted for achieving an universal abilities to reason in the real world is so enormous that the commonsense researchers always had to limit their systems. Even if the basics for commonsense processing were made 30 years ago by - among others - Minskian frames [Minsky 75], Schankian scripts [Schank 77] or casual theories of Fillmore [Fillmore 68], the realization of universal program that would behave naturally in any environment with any user was simply technically impossible. And in authors' opinion, not universal commonsensical knowledge cannot be named "commonsensical" because of it limitations. When the next decades gave us thousands times faster computers most of Artificial Intelligence proponents were sure that human-level intelligence is not so far from the achievement. Then the same researchers were alarming that it is much more difficult task than they were presuming [Fillmore 85][Minsky 88][Raskin 86] and it became obvious that the calculation power is not enough - the computer scientists understood that they still lack of data. About the same time, the biggest source of data today - the Internet - started gaining its popularity among the universities and scientific laboratories around the world. The nineties of the previous century brought much bigger hard disks what had its reflection on the corpora processing and the World Wide Web. However, from their beginning, the Internet resources were treated as a huge informational garbage dumping ground where one must use sophisticated methods to retrieve data that are useful in the trade purposes meaning. The beginning of this century showed us quite a big range of applications that are using the WWW as a corpus [Keller 03][Santamaria 03] but there is still one fundamental difference - humans retrieve information to make them richer or smarter but they might retrieve other kind of knowledge which could be helpful to make the machines smarter. The authors suggests here that knowledge which is abandoned as useless by others could become a crucial part of the "near future" applications. Because of space limits, the authors will only specify the problem in the next section and point out the topics which will be presented in more specific manner during the special session.

Contact: Language Media Laboratory, Research Group of Information Media Science and Technology, Division of Media and Network Technologies, Graduate School of Information Science and Technology, Hokkaido University, Kita-ku Kita 14 Nishi 9, 060-0814 Sapporo, Japan. TEL: (+81)(11)706-6535, FAX: (+81)(11)706-6277, {kabura,araki}@media.eng.hokudai.ac.jp

2. The Idea of General Personality

As understanding the world leads to understanding a natural language, we need as much world information as possible. The Schankian claim is that structured knowledge dominates understanding. In authors' opinion, such structures are built by occurrences - the more times something happens, the more likely it is that we will treat it as a natural phenomenon. Some of us have very different experiences than the others but there are events that are common for everyone - even if not experienced personally - known from the experiences of others. Therefore we presume that it is possible to gain the commonsensical knowledge learning only form the experiences of others. We also presume that a machine can learn the experience without experiencing them as in the case where the blind child learns about the visual world [Kielkopf 68][Fletcher 80]. The more frequent a given experience is, the more common an event can become. Hence, we assume that every human has a "general personality" which is an axis of all our behaviors. The more uncommon things we do, the farther the commonsense degree deflects from the axis which could work as, for example, a safety device, for robots. Our task is to create methods for creating such a general personality which could be a database of facts and behaviors which are common to one culture circle like a Japanese one used in our research.

3. The Basic Technical Solutions

3.1 Automatic Retrieval of Commonsense

In this part of our presentation we will briefly introduce our achievements [Rzepka 03abc] in the field of automatic commonsense retrieval from the WWW resources:

- GENTA project (introduction of Positiveness and Usualness values)
- Bacterium Lingualis method (proposition of fully automatic categorization method)

3.2 Retrieval of Personality Features

In this section of our presentation we will talk about personal scripts and their retrieval. The importance of combining stochastic and connectionistic methods with affective reasoning for creating the artificial imagination will be underlined.

3.3 Retrieval of Behavioral Features

Next, the new opportunities for retrieving data for situational and instrumental scripts will be introduced as a chance for achieving common behaviors.

4. Experiments

As the idea is fresh and the research has just begun, in thist part we will introduce the results only of the latest preliminary tests of above mentioned tasks made on two corpora created from WWW: a 10.000.000 sentences and 2.000.000 sentences sets.

5. Conclusions

5.1 Possible Usage and Perspectives

In the last part of our talk we will suggest several applications, as speech understanding and generation, translation, expert systems, QA systems or games, that could be implemented with our ideas hoping not only to raise the discussion but also to give hints for the "near future challenges". We will mention the role of emotions in computer science[Damasio 99][Minsky 04] and in our project.

5.2 Near Future Challenge

In the end, if the time lets us, we would like to spark a discussion about touch the challenges not only of the technical but also the methodical matter which becomes important for the next generations.

References

[Damasio 99] Damasio, A.R.: The feeling of what happens: Body and emotion in the making of consciousness. Harvest, New York (1999)

[Fillmore 68] Fillmore, J.C.: The Case for Case. E. Bach & R.T.Harms, eds., Universals in Linguistic Theory, New York: Holt, Rinehart & Winston, (1968) pp. 1-88

[Fillmore 85] Fillmore, C. J.: Frames and the Semantics of Understanding. In: V. Raskin (ed.), Round Table Discussion on Frame/Script Semantics, Part I, Quadernidi Semantica VI: 2, (1985) pp. 222-254

[Fletcher 80] Fletcher, J.F.: Spatial representation in blind children, Development compared to sighted children. Journal of Visual Impairment and Blindness 74 (1980) pp. 381-385

[Keller 03] Keller, F., Lapatay, M.: Using the Web to Obtain Frequencies for Unseen Bigrams. Computational Linguistics, Vol 29, No 3, September, (2003) pp. 459-484

[Kielkopf 68] Kielkopf, C.F.: The Pictures in the Head of a Man Born Blind. Philosophy and Phenomenological research, Volume 28, Issue 4, June (1968) pp. 501-513

[Minsky 75] Minsky, M.: A Framework for Representing Knowledge. In: P. H. Winston (ed.), The Psychology of Computer Vision. New York: McGraw Hill (1975) pp. 211-77

[Minsky 88] Minsky, M.: Society of Mind. Simon and Schuster, New York (1988)

[Minsky 04] Minsky, M.: The Emotion Machine (draft of Part III)
http://web.media.mit.edu/ minsky/E3/eb3.html

[Murphy 96] Murphy, R.R. : Biological and Cognitive Foundations of Intelligent Sensor Fusion. IEEE Transactions on Systems, Man, and Cybernetics, Vol. 26(1) (1996) pp. 42-51

[Raskin 86] Raskin, V.: Script-Based Semantic Theory. In: D.G. Ellis and W. A. Donohue (eds.), Contemporary Issues in Language and Discourse Processes, Hillsdale, NJ: Erlbaum, (1986) 23-61

[Rzepka 03abc] Rzepka R., Araki K., Tochinai, K.: Bacterium Lingualis - the Web-based Commonsensical Knowledge Discovery Method. Lecture Notes in Artificial Intelligence series of Springer-Verlag, volume 2843. (2003) pp. 460-467

[Rzepka 03b] Rzepka R., Araki K., Tochinai, K.: Ideas for the Web-Based Affective Processing. Proceedings of the Seventh Multi-Conference on Systemics, Cybernetics and Informatics; vol. XIV "Computer Science, Engineering Applications", pp. Orlando, Florida, USA. (2003) pp. 376-381

[Rzepka 03c] Rzepka R., Araki K., Tochinai, K.: Emotional Information Retrieval for a Dialogue Agent. "Perception and Emotions Based Reasoning" - Special Issue of "Informatica" - An International Journal of Computing and Informatics, Volume 27, Number 2, (2003) pp. 205-212

[Santamaria 03] Santamaria, C,, Gonzalo, J., Verdejo, F.: Automatic Association of Web Directories with Word Senses. Computational Linguistics, Vol 29, No 3, September (2003) pp. 485-502

[Schank 77] Schank, R., and Abelson, R. Scripts, Plans, Goals, and Understanding. Hillsdale, NJ: Erlbaum.(1977)