Bacterium Lingualis – The Web-Based Commonsensical Knowledge Discovery Method

Rafal Rzepka[1], Kenji Araki[1], and Koji Tochinai[2]

[1] Hokkaido University, Kita-ku Kita 13-jo Nishi 8-chome, 060-8628 Sapporo, Japan
{kabura,araki}@media.eng.hokudai.ac.jp,
http://sig.media.eng.hokudai.ac.jp
[2] Hokkai-Gakuen University, Toyohira-ku, Asahi-machi 4-1-40
062-8605 Sapporo, Japan
tochinai@econ.hokkai-s-u.ac.jp

Abstract. The Bacterium Lingualis is a knowledge discovery method for commonsensical reasoning based on textual WWW resources. During developing a talking agent without a domain limit, we understood that our system needs an unsupervised reinforcement learning algorithm, which could speed up the language and commonsensical knowledge discovery. In this paper we introduce our idea and the results of preliminary experiments.

1 Introduction

Numerous researchers of the last decade have underlined the importance of the relation between human emotions and our reasoning abilities [1,2,3,4], what gave birth to so called "affective computing". In our approach, the very basic feelings toward the learned elements are borrowed from humans but starting point of our method bases on a much lower level than *Homo sapiens*'. As Penrose [6] claims, the intelligence may be a fruit of our development based on Darwinian natural selection. The ideas of how to catch an animal into a trap were developed long time before a human started describing things in an abstractive manner as in logic or mathematics. Many of the artificial intelligence researchers agree that bottom-up simplified learning methods are a key to broaden the computer's capabilities and various algorithms were developed so far. The most popular ones are inspired biologically, as for example Artificial Neural Networks, genetic algorithms or insect colonies. Their weaknesses differ from one to another but they are not independent and they need laborious trainings. "Bacterium Lingualis" has a lot in common with the methods mentioned above but its differences come from the new possibilities brought by the Internet development. When we realized that pure logic is not enough for the machines to be rational and that they need all backgrounds that we have [5], it was the time to start teaching computers the commonsensical knowledge. Unfortunately it seems to be a Sisyphean task and even projects as CyC [7] or global OpenMind [8] are far away from being successful. We claim that full automatizing of this task is necessary and we should use as big corpora as possible, since, as we will demonstrate below,

G. Grieser et al. (Eds.): DS 2003, LNAI 2843, pp. 460–467, 2003.

not only the quality, but also the number of commonsensical inputs is crucial for learning the laws ruling our world. We "stepped back" in evolution and started creating an insect to begin learning from the very bottom without forcing it to behave on Cartesian philosophy. By Latin "Bacterium Lingualis" (hereafter abbreviated as BL) we mean a kind of web crawler which exploits only the textual level of WWW resources and treats it as its natural environment. We assume that cognition, by which we mean the process or result of recognizing, interpreting, judging, and reasoning, is possible without inputs other than word-level ones - as haptic or visual [9,10]. Although such data could significantly support our method, a robot which is able to travel from one place to another in order to touch something, would cost enormous amount of money, not mention a fact that current sensor technology is not ready for such an undertaking. There are several goals we want to achieve with BL. The main one is to make it search for the learning examples and learn from them unsupervisedly. For that reason we decided to move back in evolution and initiate self-developing "computational being" on the simplest level with as few human factors as possible. We assumed that all human behaviors are driven by one global reason - the pursuit of good feeling which seemed to us more adequate than simple natural selection. On the basis of above mentioned assumption we formulated "good feeling hypothesis" (hereafter abbreviated as GFH) and we implemented BL with simple negative and positive factors recognition mechanism. GFH determines the motivation for knowledge acquisition which involves language acquisition as the living environment for our program is a language itself. We imagine a language as a space where its components live together in a symbiosis. Its internal correlations are not understandable for BL and the learning task is to discover them. For exploring such an area we use simple web-mining methods inspired on Heylighen et al.'s work [11]. Most of the researches suggest that machines have to be intelligent to mine knowledge for us, we suggest that they have to mine for themselves to be intelligent.

2 Bacterium Lingualis

In order to make their idea clearer, suggest the fact of simulating basic instincts, and not to confuse their system with agents working for users, the authors decided to use a concept of an imaginary bacterium, although the rules of the language world (called here Lingua Environment) and its rules should not be considered as strictly corresponding to the biological world in which we live. BL's organism is capable of moving if the relocation is needed, to sense food and enemies, excreting what is useless. We also equipped it with enzymes and two kinds of memory, which will be detailed hereunder.

2.1 Lingua Environment

We created the BL's environment according to ideas proposed by Rzepka et al.[12]. To achieve better uniformity we decided to replace English language

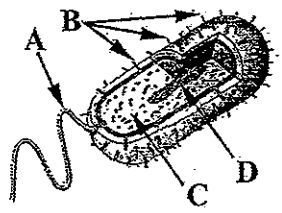

Fig. 1. Bacterium Lingualis (A – Flagellum, B – Positiveness Receptors, C – Concrete and Abstract Knowledge Memory, D – GF Cell)

homepages used in original Rzepka's work with Japanese homepages since this language seems to have an easier structure for processing especially because of its particles usage what Fillmore has suggested in his works [13]. Other reasons will be presented further in this paper. For experiments we have collected almost three millions .jp domain homepages with the Larbin robot, then after filtering off pages without sentences in Japanese and converting them into pure text files, we created a web-based raw corpus consisting of about 2.090.000 documents (approx. 20 Gb). No tagging or whatsoever was conducted.

2.2 Flagellum

Flagellum symbolizes BL's ability of movement inside its environment by which we mean text mining techniques. For these purposes we used Namazu indexing and searching system which has ability to separate words with spaces in so called wachigaki mode as Japanese sentences do not contain spaces. This helps BL to recognize what elements the contacted organism (by which we mean semantic units as text, sentence, words cluster, etc.) consists of. The morphological analysis could be done by recognizing similar patterns and statistical calculations but we assumed that omitting this level would not harm the results of BL's performance and will shorten the processing time.

2.3 Positiveness Receptors

As we mentioned before, BL is able to automatically determine its emotional reaction to the observed object. We applied a simple mechanism proposed by Rzepka [12] which calculates so called Positiveness value retrieved from the Internet users' opinions:

$$Positiveness = \frac{C_{\alpha_1} + C_{\alpha_2} * \gamma}{C_{\beta_1} + C_{\beta_2} * \gamma}$$

$$\alpha_1 = disliked, \alpha_2 = hated$$

$$\beta_1 = liked, \beta_2 = loved, \gamma = 1.3$$

Where γ is to strengthen the "love" and "hate" opinions. This method helps BL to recognize if an object is **very positive** (Positveness = 5), **positive** (P

= 4), **indifferent** (neutral) (P = 3), **negative** (P = 2) or **very negative** (P = 1), and can provide common information about what humans feel toward the given object. For instance, if the BL contacts with a single noun "beer" its reaction is positive, when the "organism" consists of two elements: "cold" and "beer", receptors send a P5 signal (very positive) to the GF cell, which will be described further. In the case of an unusual organism as a combination of "warm" and "beer", BL receives a strong negative signal. We assumed that the basic emotional information about objects is necessary for the BL's self-development in the same way as living organisms need the ability to determine what helps and what harms their development.

2.4 Particles as Enzymes

The receptors let BL contact other organisms and start symbiosis to ensure what can be learned from it. For example if BL "contacts" a noun "Sapporo" it can read its most frequent symbionts, that is, most frequent left and right neighbors of the contacted object. It is done by searching the environment together with particles characteristic to the Japanese language. Their role may be imagined as "grammar enzymes" which help to create semantic chains: "—*Sapporo*—*de*—(live, saw, take place...)", "—*Sapporo*—*ni*—(go, come, arrive...)", "—*Sapporo*—*to*—(Nagoya, compare, Otaru)", "(...known, related, belonging)—*u*—*Sapporo*—", "(...nice, nostalgic)—*i*—*Sapporo*—", "(...strange, wonderful)—*na*—*Sapporo*—". Since the causal relationships are crucial for the reasoning, several "IF enzymes" were prepared to be combined with discovered neighbors. It was relatively easy because nouns, verbs and adjectives have the same elastic if-forms in the Japanese language: "*konpyuutaa-dattara* (if computer)", "*tsukat-tara* (if to use)" or "*aokattara* (if blue)". The last example demonstrates some other interesting feature of Japanese language - many forms do not fit grammatical frames created for Indo-European languages. For instance, "kaeritai" (I want to go home) behaves identically as "nemui" (sleepy). Our further goal is to provoke BL to create its own rules of language providing it only with the basic tools, therefore "Going to Sapporo" is allowed to be treated on the same level as "cold Sapporo" if it leads to the same conclusions. This is also one of the reasons we decided to limit conventional linguistic terminology in our work and replace it with biological terms.

2.5 Concrete and Abstract Knowledge Memory (C)

BL is able to store gained knowledge. Its memory is divided into two coexisting units, Concrete Knowledge Memory and Abstract Knowledge Memory. Both are equally important but only the growth of the latter we consider as the system's growth. At this point of the system's development the concrete knowledge stands only for retrieved chains database, the abstract one describes a dictionary of automatically categorized groups of objects that frequently appear in similar combinations. This will be explained in the Method section.

2.6 GF Cell (D) and the Role of Affective Reasoning

The logicalness of human behavior is often very difficult to be analyzed with mathematic approaches. We assumed that natural language itself should decide the rules for BL system, however, it must have some inborn initial instincts as its biological equivalent. Our Good Feeling Hypothesis mentioned in the Introduction is supposed to realize this task. We presuppose that if any activity of *Homo sapiens* has been always motivated by pursuing desire of "good feeling" also the language was one of the tools for achieving this goal and is based on the same "affective logic". Therefore, "Good Feeling Hypothesis" assumes that implementing such a mechanism to a machine could help it to acquire knowledge and language. Following our thought that the GFH or defense of GFH are the reasons of every behavior, we inputed this two simple rules into GF Cell and made it default final conclusion of any reasoning while searching for different "sub-reasons" on its way. Obviously a "good feeling" varies according to the individual features but we discovered that some standards can be retrieved. Since we aim at creating unsupervised system, these standards are also supposed to play the role of safety valve. This is possible because the idea of Positiveness is based on average opinions of the homepages creators. It prevents the system from remembering chains like "killing is good" as the commonsensical facts. Another purpose of the GF Cell is to get rid of useless objects or mistaken strings that are created during the processing. This mechanism will be explained below.

2.7 Basic Method

As we mentioned earlier in this paper, in our approach we want to experiment on the lowest level of language mechanisms. Therefore, the first experiments we conducted, were to achieve automatic responses resembling Pavlovian reactions in the biological world. Such responses are needed to identify the object as pleasant, unpleasant or neutral and provoke a system's suitable behavior, by which we mean the ability of reasoning on emotional ground. On this stage, the BL uses only a very simple algorithm mostly for the associations gathering and reasons lookup. The main part which is a matter of this paper. works as follow. First, it measures the Positiveness value of a contacted object. In the beginning, it has no syntactic knowledge but by measuring the Positiveness it is able to recognize if it is analyzing a verb, a noun, an adverb or a particle etc., since the "enzymes" link only specific objects. For example a Japanese particle *DE* does not appear after a verb. Although we prefer words grouping by their connections with particles, what bases the categorization on more metaphoric grounds, for the time being we limited input only to the nouns. If, for instance, "Sapporo" was inputted, BL seeks for the most frequent input-particle strings to decide three most suitable enzymes. In the case of "Sapporo" they are *NI* (approx. 94.000 hits), *DE* (approx. 80.000 hits) and *KARA* (41.700). For better accuracy this is done by Perl API for Google. The object "Sapporo" is recorded in Concrete Knowledge Memory in the NI-DE-KARA category, which is characteristic for places. Then, the mining process starts and the neighbors of Sapporo-ni,

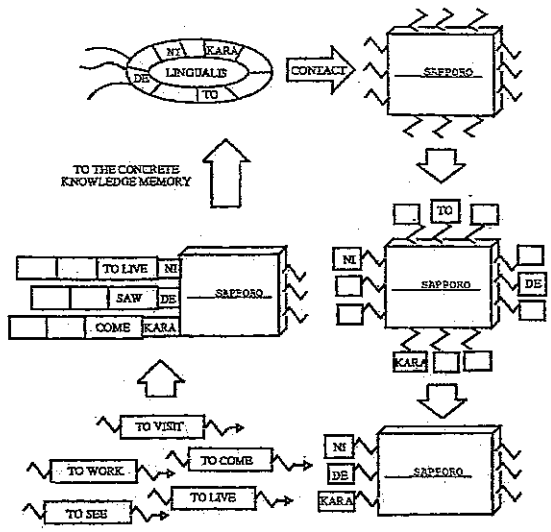

Fig. 2. The mechanism of a basic enzymatic selection

Sapporo-de and Sapporo-kara are found. Also in this case we limit the search only to the three most frequent neighbors. The candidates are taken from the first ten results and again the frequency for them is measured. For the reason that there is many mistaken retrievals and the choosing by hand would be very laborious, BL uses the Positiveness measures to eliminate mistakes as "Sapporo-dearu" where "dearu" means something different that "de aru". We could use spacing program Kakasi used in Namazu, but we try to decrease the usage of external tools to the minimum. Then, the next neighbor is searched. The process is being repeated until the last possible neighbors found. After that, the string is saved in the Concrete Knowledge Memory. If there are other objects remembered in the same category, the inputted one is replaced with every one of them:
$Sapporo$—enzyme—$string_1$—$string_2$—$string_n$
$Object_1$—enzyme—$string_1$—$string_2$—$string_n$
$Object_2$—enzyme—$string_1$—$string_2$—$string_n$
$Object_n$—enzyme—$string_1$—$string_2$—$string_n$
If one of them exists in the Lingua Environment, the abstract string is being saved at the Abstract Knowledge Memory:
$Sapporo$—enzyme—$string_1$—$string_2$—$string_n$
$Object_n$—enzyme—$string_1$—$string_2$—$string_n$
creates an abstract chain:
$NI - DE - KARA$—enzyme—$string_1$—$string_2$—$string_n$
We suppose that that collecting such abstractive rules based on the common sense may be very helpful also as a support to the other systems. The idea of using common parts in expressions to make abstractive rules is influenced by Araki et al.'s Inductive Learning [14]. If the analyzed neighbor object does not exist in the Concrete or Abstract Knowledge Memory, BL checks if it is processable with enzymes, that is if it appears with particles which determines of it is an individual object. If not, the object is deleted.

3 Experiment and Its Results

For the first test of our system, we made BL search for the connotations explaining why the object being analyzed are regarded positive or negative. A group of 10 students assigned Positiveness value for 90 words picked by BL system as those which have distinct bad or good associations. We have confirmed that 36.3% of selected words were evaluated by humans as neutral, without any emotional connotations. For proving that objects' emotional load varies from a situation, we made BL find a reasonable chain of conditions for 5 words that seemed to be indifferent for 5-7 of subjects. No word was recognized as neutral by every subject, what proves that associations of one expression are sometimes positive and other times negative depending on individual connotations. Discovering the examples of conditions or situations for both positive and negative associations was the task for the experiment. Differently from the methods proposed by Heylighen et al., BL does not only count the co-occurrences but actually mines further the inputted noun's neighbors and measures its Positiveness also if it is a verb or adjective. This done by a "'*noga* enzyme", which consists of two particles: *(V/Adj)-no-ga*. Using the same method and "*noga* enzyme", BL is able to determine that *eiga-o mi-ni iku* (to go to the movies) or *yasashii* (kind) are commonly positive and *uso-o tsuku* (to lie) or *mendoukusai* (troublesome) are distinctly negative. The words that were recognized as neutral were: *fun'iki* (mood), *dashi* (*dashi* soup), *jouken* (condition), *seikaku* (personality) and *kumiawase* (combination). Here are the retrieved pairs of reasons were: "calm atmosphere of a little bar" (+), "atmosphere of irritation before a game" (-); "dashi soup made of sea-cucumber and dried sardines" (+), "dashi soup from today [erroneous result]" (-); "conditions - new building - because it's new" (+), "conditions - changing job - to be told things" (-); "personality - cool - design - homepage" (+), "personality of myself when I can't" (-); "combination of two persons who can't eat garlic" (+), "reason not found - combination of hero and heroine" (-). We can see that there are semi-correct answers as BL still ignores some particles for output and a mistake caused by the fact, that BL's Abstract Knowledge lacks of "time" category but we think that on some stage of learning this kind of concept will be developed automatically. Output is not ready to be used by language generation programs but we think it could be used in common-sense based talking agents as Rzepka et al.'s GENTA [12] or real-life robots which have no data about newly recognized object.

4 Conclusion

We understand that the Bacterium Lingualis is in its very early stage of development but in our opinion, the initial probes seem promising and assures us that not purely connectionistic or purely stochastic methods will help to tackle the "knowledge acquisition bottleneck", but their combinations. Considering the importance of emotions, which is often neglected in AI research, we implemented BL system with a simple, automatic emotional information retrieving algorithm,

which helps not only to reason affectively, but also to automatize the verification. Developing a brain from a bacterium is certainly a difficult task but we argue against the thesis, that machines always need to simulate Aristotelian logic or learn within the borders made by our grammar rules. We believe that the feelings influence every action of a human being and the numbers of given experiences form our characters which is quite random in most situations. The Internet with its enormous WWW corpus is probably the best place where a machine can gain its experience basing only on symbols and their occurrences, also gives us many possibilities not only in the commonsensical information retrieving field, but also in other areas as developing an automatic categorization method which is one of our future works on Bacterium Lingualis.

References

1. Damasio, A.R.: Descartes' error – emotion, reason, and the human brain. Avon, New York (1994)
2. Bates, J.A.: The role of emotion in believable agents. Communications of the Association for Computing Machinery 37, ACM, pp. (1994) 122–125
3. Pinker, S.: How the Mind Works. New York, W. W. Norton (1997)
4. Picard, R.W., Klein, J.: Computers that Recognise and Respond to User Emotion: Theoretical and Practical Implications. MIT Media Lab Tech Report 538 (to appear)
5. Devlin, K.: Goodbye, Descartes. The End of Logic and the Search for a New Cosmology of the Mind. John Wiley & Sons, Inc. (1997)
6. Penrose, R.: Shadows of the Mind. A Search for the Missing Science of Consciousness. Oxford Univ. Press (1994)
7. Lenat, D.: Common Sense Knowledge Database CYC. (1995) http://www.opencyc.org/, http://www.cyc.com/
8. Stork, D.G.: "Open Mind Initiative". (1999) http://openmind.media.mit.edu/
9. Kielkopf, C.F.: The Pictures in the Head of a Man Born Blind. Philosophy and Phenomenological research, Volume 28, Issue 4, June (1968) 501–513
10. Fletcher, J.F.: Spatial representation in blind children, Development compared to sighted children. Journal of Visual Impairment and Blindness 74 (1980) 381–385
11. Heylighen, F.: Mining Associative Meanings from the Web: from word disambiguation to the global brain. Proceedings of Trends in Special Language and Language Technology, R. Temmerman (ed.) Standaard Publishers, (2001) Brussels
12. Rzepka R., Araki K., Tochinai, K.: Is It Out There? The Perspectives of Emotional Information Retrieval from the Internet Resources. Proceedings of the IASTED Artificial Intelligence and Applications Conference, ACTA Press, Malaga, (2002) pp. 22–27
13. Fillmore, J.C.: The Case for Case. E. Bach & R.T.Harms, eds., Universals in Linguistic Theory, New York: Holt, Rinehart & Winston, (1968) 1–88
14. Araki, K., Tochinai, K.: Effectiveness of Natural Language Processing Method Using Inductive Learning. Proceedings of the IASTED International Conference Artificial Intelligence and Soft Computing, ACTA Press, (2001) Cancun.