Support for Internet-Based Commonsense Processing – Causal Knowledge Discovery Using Japanese "If" Forms

Yali Ge, Rafal Rzepka, and Kenji Araki

Graduate School of Information Science and Technology, Hokkaido University
Kita-ku Kita 14-jo Nishi 9-chome, 060–0814 Sapporo, Japan
{geyali,kabura,araki}@media.eng.hokudai.ac.jp

Abstract. This paper introduces our method for causal knowledge retrieval from the Internet resources, its results and evaluation of using it in utterance creation process. Our system automatically retrieves commonsensical knowledge from the Web resources by using simple web-mining and information extraction techniques. For retrieving causal knowledge the system uses three of specific several Japanese "if" forms. From the results we can conclude that Japanese web pages indexed by a common search engine spiders are enough to discover common causal relationships and this knowledge can be used for making Human-Computer Interfaces sound more natural and interesting than while using classic methods.

Keywords: commonsense, causal knowledge discovery, human-computer interface.

1 Introduction

1.1 Need for Commonsense Retrieval

As it is easy to notice, the amount of accessible information increases with tremendous speed together with rapid growth of the Internet accessibility. More and more users write their blogs which say what Mr. Public did day by day but do not include information which other computer scientists would need for their machines to perform some task. For our approach - to make a system searching for data obvious for a human and are completely unknown for machines - these everyday tasks described in blogs are the clue. The "commonsense" retrieval never becomes an object of search queries as humans do not need to seek for such knowledge, they gather it through all their lives. However, the computers lack it and this is one of the reasons why people do not treat machines as intelligent partners, especially when it comes to conversation.

1.2 State of Art

There are several research projects coping with gathering commonsense as CyC [1] or OpenMind Commonsense [2]. CyC contains over a million hand-crafted

R. Khosla et al. (Eds.): KES 2005, LNAI 3682, pp. 950–956, 2005.

assertions and OpenMind commonsense enabled construction of a 700,000 assertion commonsense knowledge base, gathered through a web community of collaborators. But they concentrate on manual or half-automatic processing and these projects are developed only for English language. As we assume that there is too much of such knowledge to be inputted by hand, we try to make this process automatic by using simple Web-mining techniques. Our laboratory members are quite successful on achieving OpenMind results without using any human input [3][4].

2 Japanese Language Predispositions

We already have shown [5] that Japanese language has very good predispositions for text-mining for commonsense processing mostly thanks to its particles, and for that reason we concentrate on Japanese WWW resources[1]. In previous step [5] we showed that it is possible to extract simple Schankian scripts from Japanese Internet resources, this time we will prove that the Web is a vast repository for commonsensical causations which can be used for example in talking systems.

2.1 The Particles

In Japanese language there is a set of particles changing noun/pronoun's character by simple addition to the right-side of the word. For example a noun *kuruma* (a car / cars) after adding a direction indicating particle *ni* (*kuruma-ni*) suggests that following verb will be directed to the car not the opposite, automatically decreasing number of verbs candidates which could follow this noun-particle structure. One can easily build a category of verbs connected to this particular noun or a category of nouns which are glued to one particular verb by the same particle. Most popular ones particles are *wa* (Topic-Indicating), *ga* (Linking-Indicating), *no* (Possessive-Indicating), *wo* (Object-Indicating), *ni* (Direction-Indicating), *de* (Place or Means of Action-Indicating), *to* (Connective) and *mo* (Addition-Indicating). However such research does not have to be restricted to Japanese, if the similar principles could be found, an application could work with other languages – for example by using prepositions in English or regular expressions for non-gender counting in Polish.

2.2 Japanese Conditional Clauses

The causal knowledge has been a research subject but rather rarely in the perspective of being a support for commonsense processing. Papers of Sato, Kasahara and Matsuzawa al. [6] has underlined the need of commonsense processing automatization and influenced several successors. One of the most related works

[1] This also helps us to avoid cultural background mismatches as commonsense vary from country to country even if their citizens speak the same English language.

[7] concentrated on Japanese "if" form *tame* and newspaper corpus while in our research we use the WWW as the corpus and different Japanese "if" forms (*to*,*tara*,*eba*) as main query keywords (Japanese "if" forms have many useful functions - *to* is If/When After-Indicating, *tara* is If/When-Indicating, *eba* is If-Indicating, *tame* is Cause/Purpose-Indicating, *toki* is Time-Indicating and *nara* is If / Special Case-Indicating).

3 Discovering Commonsensical Causations

By using the noun keyword together with "if" forms we automatically retrieve the causal knowledge about the inputted noun from the WWW. For example, when *water* is inputted and all the forms give the same results : counting *mizu wo nomu to / mizu wo nondara / mizu wo nomeba kimochi ii* we can be quite certain that "drinking water" causes "feeling nice".

Table 1. Examples for "if bear a child"

Particle	Effect	Usualness
eba	one/something is cured	7
to	woman changes	4
tara	leave the woman	3
to	fears decrease	3
to	put on weight	3
eba	body-line gets a bit out of order	2
to	woman gets determined	2
to	woman gets stronger	2
tara	quit one's job	2
tara	(home) becomes difficult to live	2
tara	cut off from the work	1
tara	work becomes more difficult	1
eba	population will survive	1
eba	the more one has them the life gets difficult	1
eba	the more one has them one gets younger	1

3.1 Our System

Previous Module. In the beginning of our research, we decided to work with nouns as keywords for collecting minimal Schankian scripts. For this purpose we extracted relation-oriented sentences for creating dictionaries as verb dictionary, noun dictionary and n-gram dictionaries using WWW corpus (1,907,086 sentences) retrieved with Larbin robot. The verbs and nouns dictionaries consist of 79,460 verbs and 134,189 nouns retrieved with help of ChaSen [8]. For creating scripts automatically, our system had to search for the relationships between verbs and nouns and also between verb pairs. In that step, we used the verbs and nouns which had the highest occurrence (which we call "Usualness"

after Rzepka et al. [9]), as they are used by human every day and are often used in our everyday lives, for example *television*, *movie*, *food*. Also this time we experimented mostly on the daily-usage nouns.

Architecture of Current Module. Basically, the latest module for discovering commonsensical causations can be summarized into the following processing steps:

a) The user inputs a noun as a keyword;
b) The system uses our web-based corpus for frequency check to retrieve 3 most frequent verbs following the keyword noun;
c) The most frequent particle between noun keyword and 3 most frequent verbs is discovered;
d) Forms of 3 most frequent verbs are transformed into three "if" forms;
e) By using Yahoo Japan resources, the system checks if the noun-particle unit occurs with the new verb forms unit;
f) If yes - sentences which include the conditional clause are saved;
g) With help from ChaSen analyzer the system gets the most frequent commonsensical causations as following:

$$\mathbf{Ms} = \mathbf{N} + \mathbf{P}_{\max} + \mathbf{V}_{\max} + \mathbf{I} + \mathbf{S}_{\max}$$

N:*Noun keyword;*
$P_{\max}$: *the most frequent particle joining noun and verb;*
$V_{\max}$: *most frequent verb occurring after the N;*
I: *"if" form;*
$S_{\max}$: *the most frequent commonsensical causation appearing after the condition;*

4 Retrieval Experiment

As the results are very big amounts of data, for our experiment we decided to retrieve only three most frequent commonsensical causations inputting only six nouns used in everyday life (*child, cigarette, water, room, food, bath, money, mobile phone, car, light and subway*) as shown in this example:
KEY: child – if(*to/tara/eba*) –
- to bear: if/when a child is born/baby – then – ...(see Table 1)
- to raise: if/when one raises a child/baby – then – ...
- to have: if/when one has a child/baby – then – ...

In average, our system was able to discover 130 casual knowledge units (usually about 20 from retrieved 150 units was excluded because of doubling the meaning or having errors) for one verb, therefore after limiting verbs to three it was 3x130=390 units for one noun which were retrieved in approximately 110 minutes. In total it gives 6x390=2,340 units achieved. By only increasing number of noun keywords to 60 and most frequent verb limit to 10 it is easy to

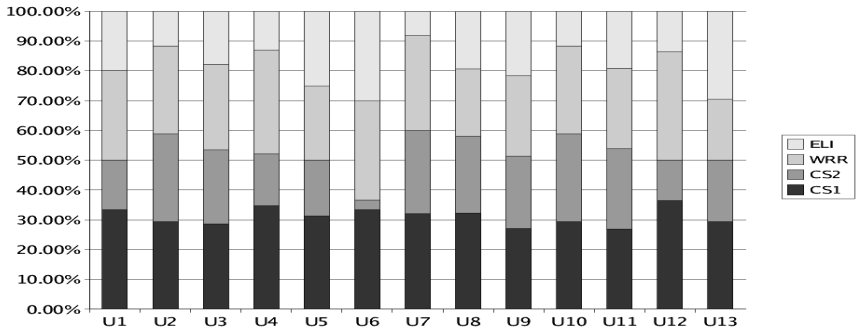

Fig. 1. Example: Degree of naturalness of utterances about a "meal" made by four systems

calculate that the system would achieve approximately 10x130x60=78,000 units which are almost two times more than Inui et al. achieved in their work [7]. We claim that in our case the units are more common as the Internet gives us a possibility to find information about usual, everyday life situations which are never described by newspapers used by Inui. For example a noun *water* gives a machine knowledge that is obvious for a human as "drinking water make body colder" or "filling with water can cause overflow".

5 Applications Experiment

In order to see user's perception of the basic commonsense knowledge included in a utterance, we performed a set of experiments basically using three kinds of utterances following input with one *keyword* from the previously mentioned set:

- ELIZA's output [ELI] (input sentence structure changing to achieve different outputs)
- WWW random retrieval output [WRR] (a shortest of 10 sentences retrieved by using *keyword* and query pattern "did you know that?")
- WWW commonsense retrieval output "high" [CS1] (sentences using common knowledge of highest usualness (most frequent mining results)
- WWW commonsense retrieval output "low" [CS2] (sentences using common knowledge of the lowest usualness (least frequent mining results).

Typical ELIZA [10] answer is "why do you want to talk about smoking" if the *keyword* is "smoking". For the same *keyword* WRR retrieved a sentence "did you know that people wearing contact lenses have well protected eyes when somebody is smoking?". An example of CS1 is "you will get fat when you quit smoking" and CS2 is "smoking may cause mouth, throat, esophagus, bladder, kidney, and pancreas cancers". We selected 10 most common noun keywords of different kinds (water, cigarettes, subway, voice, snow, room, clock, child, eye, meal) not

avoiding ones often used in Japanese idioms (voice, eye) to see if it influences the text-mining results. 13 referees were evaluating every set of four utterances in two categories – "naturalness degree" and "will of continuing a conversation degree" giving marks from 1 to 10 in both cases. The system comparison results proved that ELIZA does not eager users for continuing the chat but is still useful to keep the utterance naturalness. However, we proved that using commonsense even of the highest usualness is more natural than famous classic system (ELI 46%, CS1 54%). We also confirmed that query-based web-mining (WRR) results have slightly better user's acceptance than less common causal knowledge (CS2) which we find useful for creating a method for automatic category-based query formation depending of user's input.

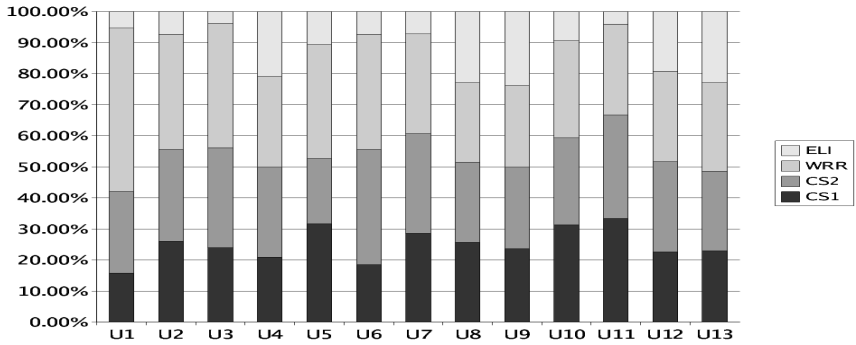

Fig. 2. Example: Degree of "continuation will" (the keyword used by four systems is "meal")

6 Conclusions and Future Work

In our first experiment, we showed how easily commonsensical causations can be discovered in enormous, mostly chaotic, data resources as WWW. There is remaining problem of time consumption but it is mostly due to the netiquette which does not allow for very fast retrieval within the search engine results. However, the commonsense processing in our future plans is supposed to work with an algorithm reducing causations by the context which will simplify query formation by increasing numbers of query keywords and making the search incomparably faster. It should also help to get rid of causation units' ambiguity, as the Internet brings also often contradictory statements like "drinking water makes you healthier" and "drinking water makes you sick". We do not have to assume that one of these claims is wrong - by discovering the contextual information we will become able to distinguish in which cases above mentioned statements are correct and in which, by contradiction, are not. In the application experiment we proved that this retrieved data can make a Human-Computer Interfaces sound more natural and interesting if we use opposite weights of commonsense

expressions. In this paper we have shown that three of several Japanese "if" forms which are *tara*, *to* and *eba* are useful for retrieving causal knowledge for commonsense processing and showed an example of such processing while creating utterances more natural than classic fully automatic methods as ELIZA which remains popular even if such approach requires laborious rules creation. We achieved higher naturalness without almost any labor and it is obvious that users prefer keep talking to systems based on the WWW that to these limited to their internal databases.

References

1. Lenat, D.: Common Sense Knowledge Database CYC, (1995) http://www.opencyc.org/, http://www.cyc.com/.
2. Singh, P.: The public acquisition of commonsense knowledge, Proceedings of AAAI Spring Symposium on Acquiring (and Using) Linguistic (and World) knowledge for Information Access. Palo Alto, CA: AAAI,(2002).
3. Skowron, M., Araki, K.: Voluntary Contributions of Unaware Internet Users? On Automatic Knowledge Retrieval from the WWW, to appear in AAAI 2005 Spring Symposium Report (Knowledge Collection from Volunteer Contributors (KCVC05), Stanford, California, March 21-23, 2005).
4. Skowron, M., Araki, K.: Automatic Knowledge Retrieval from the Web, to appear in Springer Verlag series "Advances in Soft Computing", (Intelligent Information Systems 2005. New Trends in Intelligent Information Processing and Web Mining Gdansk, Poland, June 13-16, (2005).
5. Ge, Y., Rzepka, R., Araki, K.: Automatic Scripts Retrieval and Its Possibilities for Social Sciences Support Applications To appear in Springer Verlag series "Advances in Soft Computing", (Intelligent Information Systems 2005. New Trends in Intelligent Information Processing and Web Mining Gdansk, Poland, June 13-16, (2005).
6. Sato, H., Kasahara, K., Matsuzawa, K.: Retrieval of simplified causal knowledge in text and its application. In Proc. of The IEICE, Thought and language (1998).
7. Inui,T., Inui,K., Matsumoto,Y.: What Kinds and Amounts of Causal Knowledge Can Be Acquired from Text by Using Connective Markers as Clues? DS 2003, LNAI 2843, (2003) 180–193.
8. Asahara, M., Matsumoto, Y., Extended Models and Tools for High Performance Part-of-Speech Tagger, COLING 2000,July (2000) 21–27.
9. 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.
10. Weizenbaum,J.: ELIZA: A Computer Program for the Study of Natural Language Communication between Man and Machine, Communications of the ACM,9(1): pp. 36-45, 1966.