Generality Evaluation of Automatically Generated Knowledge for the Japanese ConceptNet

Rafal Rzepka, Koichi Muramoto, and Kenji Araki

Graduate School of Information Science and Technology, Hokkaido University
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 three methods for automatic generality evaluation of commonsense statements candidates generated for Open Mind Common Sense (OMCS), which is the basis of ConceptNet, a commonsense knowledge base. By using sister terms from Japanese WordNet, our system generates new statements which are automatically evaluated by using WWW co-occurrences and hit number retrieved by a Web search engine. These values are used in three generality judgment methods we propose. Evaluation experiments show that the best of them was "exact match ratio" which achieved accuracy of 62.6% when evaluating general sentences and "co-occurrences in snippets" method scored highest with 48.6% when judging unnatural phrases. Compared to the data without noise elimination, the "exact match ratio" achieved 38.2 points increase in accuracy.

Keywords: Common Sense Knowledge, Open Mind Common Sense, ConceptNet, WordNet, Automatic Generality Evaluation.

1 Introduction

To understand language, a machine needs knowledge that human beings gather from experience since the very beginning of their lives. This knowledge is obvious and general, and we call it common sense knowledge. Many AI researchers have tried and are still trying to collect it, usually input it by hand – by specialists (as in CyC[1]) or by amateur contributors (as in OMCS[2]). Also in Japan there are engineers using general knowledge in their research, however they limit their methods to, for instance, question answering, and they create their databases manually, making it much easier to use[4]. But such limitations of usage range of knowledge is contradictory to common sense which in our opinion has more universal and inter-conceptual usage. Our approach is directed toward as fully automatic as possible methods of acquiring wide range of various kinds of knowledge people usually share. As some entries for OMCS show, the volunteers entering commonsense descriptions of the world like to joke and "generality" of many entries is doubtful ([2] states that 15% of entries do not make sense). The same

D. Wang and M. Reynolds (Eds.): AI 2011, LNAI 7106, pp. 648–657, 2011.

tendencies are visible in the latest trend - common concepts acquisition through on-line games[1]. After a while, players get bored and start to be original rather than general. However, human contributors are an important part of systems such as ConceptNet[3] based on Open Mind Common Sense where "UsedFor" or "IsA" are examples of edges which denote relationship between concepts. Although ConceptNet has been used by different researchers for a decade since MIT Media Lab has developed it, most of the projects used the English language version (and lately Chinese), while other languages versions (as Japanese) produced much smaller scientific output. The reason is quite obvious since English OMCS has currently 1,035,681 registered statements expressing common sense knowledge and there are only 14,546 for Japanese. If we could increase the number of general sentences, the usability of this knowledge would also increase. For that reason we decided to tackle this problem. Our first idea was to use WordNet[5] and WWW search to harvest Japanese concepts to acquire new commonsense statements. The basic proposal of our ideas was introduced in [6], however erroneous statements generated from Internet search gave us low accuracy not allowing the system to be somehow useful. In this paper, we propose methods to improve our system by adding automatic generality evaluation of phrases retrieved from the Web.

2 Related Work

Trials on automatic retrieval, usually based on syntactic patterns, are not new [7][8][9]. Van Durme et. al have also tried to use the WordNet in their KNEXT[10] project. Hyponym-hypernym links between noun synsets were investigated to figure out how reliably hyponyms can be viewed as mutually exclusive. Their findings (summarized on the project site [2]) were that the hypernym links were only two-thirds correspondence to true subtypes, and that the hyponyms are about 70% truly exclusive. They studied many ways to improve the extraction process, but concluded that the causes were too diverse to enable large improvement by any automated means. Hanheide et al. prove usefulness of such data presenting a similar approach for combining OMCS statements with WWW search results to quantifying commonsense knowledge for intelligent robots[11].

3 Commonsense Knowledge Generation

3.1 Definition

We define Common Sense Knowledge as an experience-based general knowledge (e.g. "dogs walk") but also broaden it to more concrete information shared by users of a given language ("Todai-ji is a temple in Japan", "Madonna sings", "you can work at Sony". Such broader definition increases capability of non-task oriented dialog systems which we are also working on.

[1] http://nadia.jp

[2] http://www.cs.rochester.edu/ schubert/projects/world-knowledge-mining.html

3.2 System Overview

The idea is to use existing OMCS sentences and exchange nouns with sister terms from the WordNet dictionary to generate new similar statements and then use a Web search to determine how usual the generated knowledge is. Figure 1 provides an overview of our system. By "sister terms" we mean hyponyms under the same hypernyms. For example, "lions roar" can be transformed into "tigers roar". Then, to remove possible noise (untrue or unnatural statements) such a phrase becomes a query for search engine, which in this study is Yahoo! Japan[3]. Usualness (generality) calculation uses thresholds which will be described later in detail.

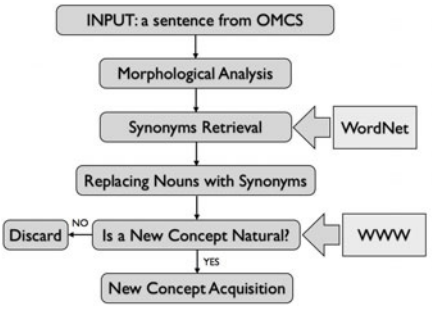

Fig. 1. Overview of our system for harvesting concepts by using WordNet sister terms and evaluating them by WWW search

3.3 Japanese WordNet

WordNet, developed at Princeton University, is a semantic lexicon consisting of concepts called synsets. Words that are similar are kept within the same synset. A synset is labeled as "number ID - part of speech" where, for instance, "n" means a "noun", and "v" indicates a verb. It is also connected to other synsets associated by relationships like hyponymy, hyperonymy or meronymy. In Japanese WordNet there are 57,238 synsets, 93,834 words and 158,058 pairs of synsets and words.

3.4 Retrieving Sister Terms and Generating Sentences

A sentence from OMCS set for Japanese language becomes an input to our system. Then its noun is replaced by a sister term from the WordNet, so, for example, "(one can) throw a ball" produces statements like "(one can) throw a fastball", "(one can) throw a Frisbee" or "(one can) throw a [playground] slide". As you can see from the last example, statements generated by nouns from a broad category as "toys" will not always produce a general, commonsense

[3] `http://search.yahoo.co.jp`

knowledge and such erroneous phrases cannot be added to the knowledge base. Therefore the noise elimination becomes crucial for newly generated data quality – in the next section we explain in details what methods we developed.

4 Noise Elimination Methods

To eliminate semantically erroneous generations we propose three shallow web-mining methods, which we named "co-occurrences in snippets", "exact match ratio" and "conjugated keywords hit ratio".

4.1 Co-occurrences in Snippets (a)

Verbs, nouns and (if they appear) adjectives are extracted from the input sentence, and the original particle[4] is used to form a search query "$NounParticle$" + "$(Verb|Adjective)$". Web search using such a query outputs set of snippets (short summary passages output by a search engine) and the system counts how many times both queried phrases occurred. The condition is to be in the same sentence and in the same order and this type of results we call "co-occurrences in snippets". We define "sentence" here as a phrase between punctuation marks as dots, commas, exclamation marks, question marks, etc. We set a threshold for co-occurrences in snippets, and if their number falls below the threshold then queried sentence is determined as noise. Thresholds are explained later in the paper.

4.2 Exact Match Ratio (b)

Unlike the "co-occurrences in snippets" method, here noun, particle and verb (or adjective) create one exact match query (without OR operator): "$NounParticle$ $(Verb|Adjective)$". At the same time following additional queries are created: "$NounParticle$" + "$Verb|Adjective$", "$NounParticle$", and "$Verb|Adjective$". System uses search engine results for all these queries to calculate an "exact match ratio" with the Formula (1). Again thresholds are set to eliminate erroneous output.

$$Pp = \frac{N_p}{N_n + N_v - N_c} \tag{1}$$

- P_p: exact match ratio
- N_p number of hits for "$NounParticle(Verb|Adjective)$"
- N_n: number of hits for "$NounParticle$"
- N_v: number of hits for "$Verb|Adjective$"
- N_c: number of hits for "$NounParticle$" + "$Verb|Adjective$"

[4] Japanese particles are suffixes that immediately follow the modified noun, verb, adjective, or sentence. For example in *booru o nageru* (throw a ball, to throw a ball, throwing a ball, one throws a ball, etc.) *o* states that the noun it follows is a direct object of the action described by following verb.

4.3 Conjugated Keywords Hit Ratio

In this method we decided to add a natural language processing module for stemming as search engines ignore the fact that verbs conjugate. The main reason for adding this technique is to increase number of hits, which allows to get a better accuracy of the investigated data. Phrase "eat a cake" after stemming can find five or six other forms which may be Japanese equivalents of "eating a cake","ate a cake", '"will eat" or "will be eating"[5]. Calculations are similar (it is the sum of all stemmed keywords) to the previous method (see Formula (2)) and also here adequate thresholds are set.

$$Pc = \sum \frac{N_p}{N_n + N_v - N_c} \tag{2}$$

P_c: conjugated keywords hit ratio (see Formula 1 for the full description).

Table 1. Results of threshold setting experiment

(a) "co-occurrence in snippets"

Authors' evaluation	Number of Sentences	Average Appearance in Snippets
0 points	545	7.3
1point	211	11.6
2points	244	16.4

(b) "exact match ratio"

Authors' evaluation	Number of Sentences	Average Ratio of Exact Matching
0 points	723	0.00185
1point	56	0.00390
2points	221	0.00414

(c) "conjugated keywords hits ratio"

Authors' evaluation	Number of Sentences	Average Ratio of Conjugated Keywords
0 points	709	0.000106
1point	60	0.000587
2points	231	0.00131

5 Preliminary Experiments for Setting Noise Elimination Thresholds

As mentioned in previous sections it was necessary to set thresholds to eliminate as much unnatural output as possible. Fifty sentences including nouns were

[5] They cover more than tenses but the examples show only this type for the sake of simplicity.

randomly selected from OMCS Japanese data and system used sister terms to harvest candidates. It produced 13,240 sentences and we randomly selected 1,000 of them, and then a manual evaluation was performed by a native speaker of Japanese. The following criteria were used in the evaluation: "unnatural knowledge = 0 points", "possible but not general knowledge = 1 point", "general knowledge = 2 points". Table 1 shows results for all three methods described in Section 4. In case of "co-occurrence in snippets" (a), more than half of the generated sentences appeared to be unnatural , while about 24% of the acquired phrases was evaluated as useful general knowledge. "Exact match ratio" and "conjugated keywords hit ratio" produced 22% and 23% common sense statements respectively. Accordingly to these results we have decided that in case of (a), threshold for unnatural sentences is less than 7 co-occurrences of queried phrases in snippets, for non-general is more than 7 and less than 11, and for general there must be more than 11. Scores for (b) were set to 0.00185, 0.00390, and 0.00414; while for (c) we set number of hits threshold: 775,849 as the unnaturalness borderline, 860,909 for "arguable zone" and 1,349,698 as a starting point for regarding outputs as natural.

Table 2. Automatic vs. manual evaluation ("co-occurrence in snippets")

System Evaluation Score	Evaluators' Score	Average Number of Sentences
2 points	2 points	27.5
	1 point	12.5
	0 points	10.0
	Ratio of Correct Answers	**55.0%**
1 point	2 points	21.0
	1 point	14.0
	0 points	15.0
	Ratio of Correct Answers	**28.0%**
0 points	2 points	14.7
	1 point	11.0
	0 points	24.3
	Ratio of Correct Answers	**48.6%**

6 Evaluation Experiment and Its Results

After setting thresholds described in the previous section, we have performed experiments in order to see how accurately our system eliminated noisy, non-general knowledge from harvested data and how confident it can be about correct output. The rating method was the same as in the preliminary experiment and 50 statements (after noise elimination) for each method were randomly chosen (150 sentences in total). The same sets were also evaluated (in the same 3 grade scale) by 6 subjects who were two male college students from the science department plus two male and two female students from the literature department.

Table 3. Automatic vs. manual evaluation ("ratio of exact matches")

System Evaluation Score	Evaluators' Score	Average Number of Sentences
2 points	2 points	31.3
	1 point	10.9
	0 points	7.8
	Ratio of Correct Answers	**62.6%**
0 points	2 points	25.5
	1 point	13.8
	0 points	10.7
	Ratio of Correct Answers	**21.4%**

Table 4. Automatic vs. manual evaluation ("conjugated keywords hit ratio")

System Evaluation Score	Evaluators' Score	Average Number of Sentences
2 points	2 points	31.0
	1 point	9.5
	0 points	9.5
	Ratio of Correct Answers	**62.0%**
0 points	2 points	28.3
	1 point	10.2
	0 points	11.5
	Ratio of Correct Answers	**23.0%**

Table 5. Evaluators agreement ("co-occurrence in snippets")

System Evaluation Score	Evaluators' Score	3 Evaluators	4 & More Evaluators
2 points	2 points	25	25
	1 point	5	5
	0 points	10	5
	Ratio of Correct Answers	**62.5%**	**71.4%**
1 point	2 points	22	20
	1 point	7	5
	0 points	15	13
	Ratio of Correct Answers	**15.9%**	**13.2%**
0 points	2 points	12	8
	1 point	8	2
	0 points	25	21
	Ratio of Correct Answers	**55.6%**	**67.7%**

The experimental results for method (a) are shown in Table 2. In 55.0% of the cases, system correctly estimated that knowledge is general, in 28.0% of the cases that it is non-general and in 48.6% that it was unnatural and should be discarded. As defining what is general and what is not is often difficult even for human evaluators, we also took into account the agreement between users.

Table 6. Evaluators agreement ("exact match ratio")

System Evaluation Score	Evaluators' Evaluation	3 Evaluators	4 & More Evaluators
2 points	2 points	36	31
	1 point	6	4
	0 points	7	6
	Ratio of Correct Answers	**73.5%**	**75.6%**
0 points	2 points	28	21
	1 point	8	6
	0 points	7	6
	Ratio of Correct Answers	**16.3%**	**18.2%**

Table 7. Evaluators agreement ("conjugated keywords hit ratio")

System Evaluation Score	Evaluators' Evaluation	3 Evaluators	4 & More Evaluators
2 points	2 points	35	33
	1 point	4	2
	0 points	7	7
	Ratio of Correct Answers	**76.1%**	**78.6%**
0 points	2 points	29	24
	1 point	8	7
	0 points	9	7
	Ratio of Correct Answers	**19.6%**	**22.6%**

Table 5 shows that in cases of less arguable knowledge (0 and 2 points, more than 4 evaluators agreed), the system's accuracy increases from 55.0% to 71.4% (general knowledge) and from 48.6% to 67.7% (unnatural knowledge). Because of this lack of agreement and the fact that system discovered too few[6] sentences that could be evaluated as not general, we decided to exclude it from the evaluation process. Tables 3 and 4 show experimental results for methods (b) and (c), Tables 6 and 7 indicate results where user agreement is considered. In case of "exact match ratio", a significant increase of accuracy (62.6% to 75.6%) can be observed for general knowledge but in discovering unnatural statements this method appeared worse (decreased from 21.4% to 18.2% when agreed by more than 4 evaluators). "Conjugated keywords hit ratio" method performed much better in case of common sense statements (62.0% to 78.6%) but again was slightly worse in discovering erroneous knowledge (23.0% to 22.6%). As shown in Table 1(a), without noise elimination, we could retrieve only 24.4% of usable general knowledge. Method (a) "Co-occurrence in snippets", after eliminating erroneous statements, allowed to correctly find 55.0% of such knowledge. In case of "exact match ratio" (b), 62.6% of the generations were correct and of "conjugated keywords hit ratio" (c), 62.0% were evaluated as proper automatic judgment. The highest accuracy was achieved by method (b) - compared to the results without noise removal there was 38.2 points improvement in accuracy.

[6] Too few to be statistically significant.

There were 7 sentences which were evaluated "0 points" by the system and "2 points" by more than 4 evaluators. Five of these statements were generated by the morphological analysis tool, which cuts off suffixes that are nouns but have different meaning when used separately. For example *-hen* is used as a "compilation suffix"; when added to novels or poems means "collection of novels" or "collection of poems", but by itself it sounds odd. As we decided to use one noun, not a noun phrase, this type of errors depending on third party tools was inevitable. Another problem was context dependency – one of the sentences that showed a significant difference in evaluation was "summer is cold". Depending on places and particular days, summers can be cold cold and such statements are not rare on the WWW.

7 Conclusions and Future Work

In this paper we introduced three methods for automatic generality evaluation of Japanese sentence candidates generated for Open Mind Common Sense (OMCS), which is the base for ConceptNet, a freely available commonsense knowledge base and NLP tool-kit developed by MIT. By using sister terms from Japanese WordNet, our system was able to generate new statements that possibly represent common sense knowledge, however only part of newly produced outputs are obviously general. Therefore we implemented a module using Yahoo! Japan search engine to retrieve co-occurrences and hit numbers, which became a base for three methods we proposed. Evaluation experiments showed that the best of them was "exact match ratio" method which achieved accuracy of 62.6% when evaluating general sentences. For judging unnatural (impossible) knowledge, "co-occurrences in snippets" method scored highest with 48.6%.

As we noticed that human contributors get bored soon after starting to type commonsense statements, we assume it would be much faster and efficient to let them choose if something indicates general knowledge or not. Using our methods would definitely decrease burden of the proper entry choice task by showing only statements which scored 2 points in the 0-1-2 scale of generality to an evaluator. However, to come closer to accuracy allowing fully automatic generation, there is still plenty of room for future work. During the development and experiments we noticed many tendencies that could allow improvements. The more examples are found, the wider coverage we could get. There is thus a need for extending queries, for example by alternating particles – Japanese topic indicating particle *wa* can be replaced with subject indicating particle *ga*. We will also add techniques for so called (in linguistics) "genericity" and use grammatical structures and words that often suggest generality of a sentence (e.g. adverbs like "usually"). We also noticed that context dependent errors can be reused with negations to find new knowledge and every arguable statement could be rewritten and processed again. Combinations of "usually", "not" and "but" could also bring interesting results, therefore we want to increase quality by widening the web-mining process by taking grammatical information and neighboring

words (also noun phrases) into consideration. We are also planning to transfer proposed shallow methods to ConceptNet versions for other languages that suffer the same lack of OMCS sentences as Japanese.

References

1. Lenat, D., et al.: Common Sense Knowledge Database CYC (1995)
2. Singh, P., Lin, T., Mueller, E.T., Lim, G., Perkins, T., Zhu, W.: Open Mind Common Sense: Knowledge Acquisition from the General Public. In: Meersman, R., et al. (eds.) CoopIS 2002, DOA 2002, and ODBASE 2002. LNCS, vol. 2519, pp. 1223–1237. Springer, Heidelberg (2002)
3. 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)
4. Oe, N., Watabe, H., Kawaoka, T.: The construction method of commonsense judgment system for understanding the conversation of geography (in Japanese). IPSJ SIG Notes. ICS (24), 163–168 (2005)
5. Fellbaum, C.: WordNet: An Electronic Lexical Database. MIT Press, Cambridge (1998)
6. Muramoto, K., Rzepka, R., Araki, K.: Generation method of sentences for common sense knowledge bases using WordNet and web search (in Japanese). Kotoba Kenkyuu-kai SIG Technical Report 35, 1–7 (2010)
7. Chklovski, T.: Learner: A system for acquiring commonsense knowledge by analogy. In: K-CAP 2003, pp. 4–12 (2003)
8. Clark, P., Harrison, P.: Large-scale extraction and use of knowledge from text. In: K-CAP 2009, pp. 153–160 (2009)
9. Yu, C., Chen, H.: Commonsense Knowledge Mining from the Web. In: Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence (AAAI 2010), pp. 1480–1485 (2010)
10. Van Durme, B., Michalak, P., Schubert, L.K.: Deriving generalized knowledge from corpora using WordNet abstraction. In: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, pp. 808–816 (2009)
11. Hanheide, M., Hawes, N., Gretton, C., Aydemir, A., Zender, H., Pronobis, A., Wyatt, J., Gobelbecker, M.: Exploiting probabilistic knowledge under uncertain sensing for efficient robot behaviour. In: Proc. of IJCAI 2011 (2011)