Automatic Evaluation of Commonsense Knowledge for Refining Japanese ConceptNet

Seiya Shudo and **Rafal Rzepka** and **Kenji Araki**
Graduate School of Information and Technology, Hokkaido University
Sapporo, Kita-ku, Kita 14 Nishi 9, 060-0814, Japan
{shudo,rzepka,araki}@ist.hokudai.ac.jp

Abstract

In this paper we present two methods for automatic common sense knowledge evaluation for Japanese entries in ConceptNet ontology. Our proposed methods utilize text-mining approach, which is inspired by related research for evaluation of generality on natural sentences using commercial search engines and simpler input: one with relation clue words and WordNet synonyms, and one without. Both methods were tested with a blog corpus. The system based on our proposed methods reached relatively high precision score for three relations (*MadeOf*, *UsedFor*, *AtLocation*). We analyze errors and discuss problems of common sense evaluation, both manual and automatic and propose ideas for further improvements.

1 Introduction

The lack of commonsense knowledge has been one of main problems for creating human level intelligent systems and for improving their tasks as natural language understanding, computer vision, or robot manipulation.

Researchers have tackled with this deficiency usually taking one of the following three approaches. One is to hire knowledge specialists to enter the knowledge manually and CyC (Lenat, 1995) is the most widely known project of this kind. Second is to use crowdsourcing. In Open Mind Common Sense project (OMCS) (Singh et al., 2002), non-specialists input phrases or words manually, which generates knowledge in relatively short time. For making the input process less monotonous, researchers also use Games With A Purpose (GWAPs), for instance *Nāja-to nazo nazo*[1] (Riddles with Nadya)[2] for acquiring Japanese commonsense knowledge. Third approach is to use text-mining techniques. KNEXT (Schubert, 2002), NELL[3] or WebChild (Tandon et al., 2014) are famous projects for acquiring commonsense knowledge automatically.

Last two approaches are immune to quality problems. For example, knowledge acquired through Nadya interface reached 58% precision (Nakahara and Yamada, 2011), and NELL system reached 74% precision (Carlson et al., 2010). This is because public contributors input and source Web texts tend to be noisy. Therefore, acquired knowledge should be evaluated, but there is no gold standard method for estimating whether acquired knowledge is commonsensical or not. Usually, manual evaluation by specialists or by crowdsourcing (Gordon et al., 2010) is used. However, this is costly and time-consuming, and even specialists have different opinions on concepts' usualness. Another method is to evaluate automatically acquired knowledge by utilizing it in some tasks. For example, there is a research using IQ tests (Ohlsson et al., 2012) for commonsense knowledge level estimation, but it does not help improving or refining quality of existing or newly acquired concepts.

In this paper, we present automatic evaluation system for commonsense knowledge. Our approach is to use frequency of phrase occurrences in a Web corpus. There is a previous research using Internet resources and Japanese WordNet (Bond et al., 2009) for evaluating generality of natural sentences from

[1] Original Japanese words are represented in italic throughout the paper.

[2] http://nadya.jp/

[3] http://rtw.ml.cmu.edu/rtw/

Proceedings of the 12th Workshop on Asian Language Resources,
pages 105–112, Osaka, Japan, December 12 2016.

OMCS (Rzepka et al., 2011). In that research, frequency of occurrence in Yahoo Japan search engine[4] search results snippets are used to determine thresholds for eliminating noise and verb conjugation is used to increase number of hits. Our approach for evaluating commonsense knowledge is similar but we aim at higher precision without using commercial search engines. Currently access to commercial engines is limited even for a researchers so we decide to introduce methods that can be used also with relatively smaller, self-made (crawled), corpora. Our research can also improve crowdsourcing methods, because it can decrease costs or be less time-consuming if distinctly wrong entries are automatically filtered out. Last but not least, we work on concepts and relations while in previous research only simple word pairs (e.g. "to throw" + "a ball") were used.

Our contributions presented in this paper can be summarized as follows:

- We evaluate Japanese commonsense knowledge from ConceptNet (Speer and Havasi, 2012) (explained in the next section) by using phrase occurrences in a blog corpus.
- We apply proposed methods to three relation types to investigate their flexibility.
- We analyze evaluation errors, discuss problems of our methods and propose their expansion for increasing efficiency of automatic evaluation.

2 Japanese ConceptNet

ConceptNet is a semantic network-like ontology allowing to process commonsense knowledge. It is created from other sources as hand-crafted OMCS or GlobalMind (Chung, 2006), JMdict[5], Wiktionary[6] and so on. In ConceptNet, there are two ways of representation. First is a graph structure where nodes show concepts, and their relations such as "IsA" or "PartOf". One set of two concepts and their relation is called an assertion. This is represented by $Relation(Concept1, Concept2)$ abbreviated from now as $(\mathcal{C}_1, \mathcal{R}, \mathcal{C}_2)$. Another way of representation is a natural sentence, and there are entries in various languages as English, Chinese, German, Korean, Portuguese and also Japanese. In Japanese ConceptNet concept terms are in Japanese, but relations are in English (the same is true for all non-English languages). For this research we used latest version 5.4[7]. Japanese ConceptNet contains 1.08 million assertions in total, but more than 80% of them belong to "TranslationOf" relation, therefore we treated them as irrelevant to the commonness evaluation task.

For this research we chose three relations for the first series of trials: "MadeOf" (1008 assertions), "UsedFor" (2414 assertions), and "AtLocation" (13213 assertions). Main reason for choosing these relations is that they can be distinctly associated with physical objects, while e.g. "RelatedTo" relation (98.6 thousands assertions) is very often semantically vague and needs different approach for evaluating its correctness.

3 System Overview

In this section we present an outline of our system for automatic commonness estimation of ConceptNet assertions (see Figure 1). In the first step, our system searches a blog corpus (Ptaszynski et al., 2012) for left $\mathcal{C}_1$ and right $\mathcal{C}_2$ concepts, and then parses snippets of search results and concepts using morphological analyzer MeCab[8]. Without this process, if an assertion shows that one concept includes the other concept such as ($\mathcal{C}_1$) *karē* (curry), ($\mathcal{R}$) "MadeOf", and ($\mathcal{C}_2$) *karēko* (curry powder), ($\mathcal{C}_2$) *karēko* end up also matching as ($\mathcal{C}_1$) *karē*.

Concepts can be represented in multiple morphemes including not only nouns but also verbs, adjectives or particles. If there are compound nouns in a concept, system treats them as one noun. In the next step, our system checks whether each sentence contains a relation clue word or not. We manually selected

[4] http://nadya.jp/

[5] http://www.edrdg.org/jmdict/j_jmdict.html

[6] https://en.wiktionary.org/wiki/Wiktionary:Main_Page

[7] http://conceptnet5.media.mit.edu/downloads/current/conceptnet5_flat_json_5.4.tar.bz2

[8] http://taku910.github.io/mecab/

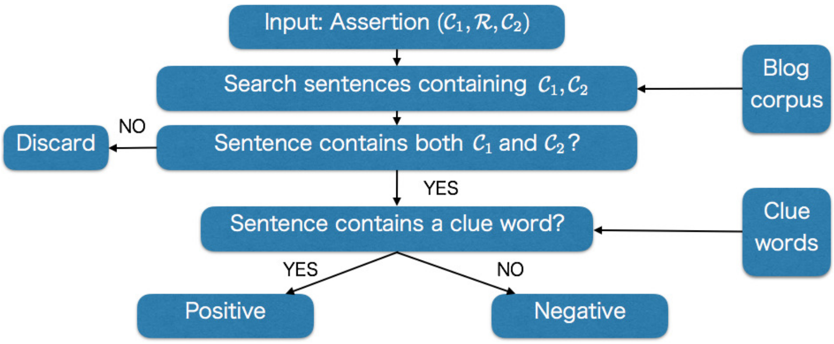
Input: Assertion ($\mathcal{C}_1$, $\mathcal{R}$, $\mathcal{C}_2$)
Search sentences containing $\mathcal{C}_1$, $\mathcal{C}_2$
Blog corpus
Discard
NO
Sentence contains both $\mathcal{C}_1$ and $\mathcal{C}_2$?
YES
Sentence contains a clue word?
Clue words
YES
NO
Positive
Negative

Figure 1: Overview of our system for evaluating assertion by using a blog corpus and clue words.

clue words in Japanese semantically related to a given relation ($\mathcal{R}$) for retrieving their co-occurrences with concepts.

For evaluating “MadeOf” assertion, we used *tsukuru* (to make). For “UsedFor” assertion, we chose *tsukau* (to use). Because these basic verbs do not ensure sufficient number of hits, we added their synonyms from Japanese WordNet (Bond et al., 2009). It is a lexical database consisting of synsets which are represented in nouns, verbs, adjectives, and adverbs. One word can be linked to multiple meanings and even more synonyms. For instance, *tsukuru* (to make) has 21 synsets which provide 111 synonyms in total. Some of them are rare or semantically far from the basic clue verb. For this reason we chose only 10 synonyms with the following procedure. First, we extracted synonyms used in two or more synsets linked to a given clue word (relation verb), and then selected 10 synonyms with the highest frequency in the blog corpus. To increase hit number even further, we conjugated all verbs, which gave up to 7 forms depending on the verb type. For instance, except basic *tsukau* (to use) following forms were also used as queries: *tsukawa-*, *tsukao-*, *tsukae-*, *tsukat-*, *tsukai-*.

To investigate differences between precision and recall we introduced two separate methods with different matching conditions. In order to evaluate an assertion, the most natural approach would be to match $\mathcal{C}_1$, $\mathcal{R}$, and $\mathcal{C}_2$ in one sentence, e.g. “butter ($\mathcal{C}_1$) is **made** ($\mathcal{R}$) from milk ($\mathcal{C}_2$)”. Therefore, in our first proposed method all these three elements must occur at least once in one sentence (we call it a “All Elements” method). Because this method is expected to achieve rather low recall, we also proposed a second method requiring only $\mathcal{C}_1$ and $\mathcal{C}_2$ to co-occur in one sentence (“Concepts Only” method). For “AtLocation” relation we selected two clue verbs with connotations of existence: "*aru*" for animate and "*iru*" for inanimate nouns. Although both verbs are widely used, "*aru*" and "*iru*" cause significant amount of noise because they are also used as auxiliary verbs, e.g. *tabete-**iru*** (eat**ing**). Therefore, for “AtLocation” assertions we altered the second method used for “MadeOf” or “UsedFor” by replacing relations $\mathcal{R}$ with place-indicating particles: “$\mathcal{C}_2$ - *ni* $\mathcal{C}_1$” and “$\mathcal{C}_2$ - *de* $\mathcal{C}_1$”. *Ni* and *de* convey a preposition function similar to “in” or “at” in English.

4 Experiments and Results

To confirm the efficiency of our proposed system in automatic evaluating commonness of a concept, we performed series of experiments. From ConceptNet 5.4 we randomly selected 100 assertions for each of the three relations under investigation. To create the correct data set, 10 annotators (one female student, 8 male students, one male worker, all in their 20’s) evaluated 300 assertions presented in Japanese sentences. We needed to manually create these using a fixed template, because there were many cases where ConceptNet did not contain a natural sentence in Japanese, and the way of expression was not united. For instance, in case of ($\mathcal{C}_1$) *banira* (vanilla), ($\mathcal{R}$) “MadeOf”, and ($\mathcal{C}_2$) *gyūnyū* (milk), we inserted all elements into following template: “*Banira-wa gyūnyū-kara tsukurareru*” (vanilla is made from milk). As we treated unarguably common facts starting zero point with growing peculiarity of assertinos, annotators evaluated commonness of such sentences using 10 points scale (from 1 to 10, where 1 is common sense, and 10 is non-common sense). We treated the results labelled 1-5 as usual (commonsensical,

Table 1: Possible False / True relations between human and automatic evaluation.

	System Positive	System Negative
Questionnaire True	TP	TN
Questionnaire False	FP	FN

$$
\begin{aligned}
Accuracy &= \frac{TP + FN}{TP + TN + FP + FN} && (1)\\
Precision &= \frac{TP}{TP + FP} && (2)\\
Recall &= \frac{TP}{TP + TN} && (3)\\
F - score &= \frac{2 \times Precision \times Recall}{Precision + Recall} && (4)
\end{aligned}
$$

Figure 2: Equation for calculating f-score.

correct), and 6-10 as unusual (not commonsensical, incorrect).

In Table 1 and Figure 2, we show possible combinations of relations between human annotators and system agreement, and f-score calculation equation. Experiments results showed that our proposed methods achieved high precision for each type of relation (see Tables 2, 3, and 4). These results also proved that the proposed text-mining approach can be used to evaluate relational commonsense knowledge without commercial search engines and thresholds manipulation.

5 Error Analysis

We analysed errors of "All Elements" method ($\mathcal{C}_1$, $\mathcal{R}$, $\mathcal{C}_2$) by reading source sentences which caused incorrect commonness estimations and by comparing system's results with human annotations. It appears that annotators' evaluation scores differ significantly: only three assertions out from 300 were the same (all three judged them as false). For example, *Kogata reizōko-niwa hoteru-ga aru* (There is a hotel in a small refrigerator) and *Tokyo-niwa Fujisan-ga aru* (There is Mt. Fuji in Tokyo) were evaluated as explicitly incorrect. Very small number of agreed evaluations shows clearly the difficulty with making an evaluation system for commonsense knowledge due to discrepancies in human annotators opinions.

Below, we present examples explaining reasons for erroneous automatic evaluations. There are some

Table 2: Evaluation results for "MadeOf" relations ("All Elements" and "Concepts Only" methods).

MadeOf	Accuracy	Precision	Recall	F-score
All Elements Method ($\mathcal{C}_1, \mathcal{R}, \mathcal{C}_2$)	0.450	0.780	0.410	0.538
Concepts Only Method ($\mathcal{C}_1, \mathcal{C}_2$)	0.640	**0.792**	0.730	0.760

Table 3: Evaluation results for "UsedFor" relations ("All Elements" and "Concepts Only" methods).

UsedFor	Accuracy	Precision	Recall	F-score
All Elements Method ($\mathcal{C}_1, \mathcal{R}, \mathcal{C}_2$)	0.530	**1.00**	0.413	0.584
Concepts Only Method ($\mathcal{C}_1, \mathcal{C}_2$)	0.650	0.868	0.662	0.735

Table 4: Evaluation results for "AtLocation" relations ("All Elements" and "Concepts Only" methods).

AtLocation	Accuracy	Precision	Recall	F-score
All Elements Method ($\mathcal{C}_1, \mathcal{R}, \mathcal{C}_2$)	0.500	**0.615**	0.285	0.390
Concepts Only Method ($\mathcal{C}_1, \mathcal{C}_2$)	0.550	0.582	0.696	0.634

cases where, an assertion was judged as non-commonsense knowledge but sentences in the corpus suggested otherwise. For instance, ($\mathcal{C}_1$) *furaipan* (frying pan), ($\mathcal{R}$) "MadeOf", and ($\mathcal{C}_2$) *arumi* (aluminum) elements were discovered in the sentence *Wagaya-wa moppara daisō-de utteiru saisho-kara arumi-no furaipan-to setto-ni natteiru-node tukutte-masu* (In my house we always make it because it is in a set with an aluminum frying pan sold by Daisō). The system matched "make" as "we make" (meaning "to cook"), but it should be related to "aluminum frying pan" (meaning "frying pan made of aluminum").

Another problem arises from the fact that some concepts in ConceptNet are not written in simple, commonsensical manner or are simply strange. For example, for ($\mathcal{R}$) "MadeOf" we have ($\mathcal{C}_1$) *aisatsu* (greeting) and ($\mathcal{C}_2$) *genshō-kara yomitorareru imi* (meaning that can be read from a phenomenon). The reason is that in knowledge gathering systems like Nadya or GlobalMind some contributors try to be original. It is difficult to remove all inappropriate assertions by knowledge providers, so they end up remaining in the database. Annotators judged the given assertion above as non-commonsense knowledge. However, in some cases such as ($\mathcal{C}_1$) *esukarēta* (escalator), ($\mathcal{R}$) "UsedFor", and ($\mathcal{C}_2$) *ue-ni agattari oritari suru* (to go up and down) or ($\mathcal{C}_1$) *henken* (prejudice), ($\mathcal{R}$) "MadeOf", and ($\mathcal{C}_2$) *taishō-ni tai-suru jōhō-no ketsujo* (information shortage of object information) were judged as common sense. From these results we can conclude that contributors provided semantically correct knowledge, although their input was unorthodox on the lexical level.

Evaluation also seems to depend on how the assertion was presented in the questionnaire. For assertions like ($\mathcal{C}_1$) *eiga* (movie), ($\mathcal{R}$) "UsedFor", and ($\mathcal{C}_2$) *kanshō-suru* (to watch), it would be more natural to say *eiga-wo kanshō-suru* (to watch a movie) than *kanshō-suru-niwa eiga-wo tsukau* (for watching a movie is used) which we created to keep all forms consistent. *Kanshō-suru* (to watch) implicitly indicates *tsukau* (to use), therefore it is difficult to create a natural sentence in such cases without allowing synonyms or more specific verbs.

Different problems were caused by the fact that the proposed system did not use part of speech information during the matching processing. This led to ambiguity which is visible in an example of the following assertion: ($\mathcal{C}_1$) *sutorēto* (undiluted), ($\mathcal{R}$) "MadeOf", and ($\mathcal{C}_2$) *arukōru* (alcohol). *Sutorēto* has two meanings: "directly"/ "frankly" and "undiluted". While it was correctly evaluated as uncommon by majority of evaluators, the system labelled "alcohol is made of straight" as common. This is because the following corpus sentence was retrieved and used for matching: *Shōsei-mo kyō byōin-no kensa-de, shinitaku-nai nara, kyō-kara arukōru-wo tate-to storēto-ni iwaremashita* (At the hospital, I was also **told straight** that if I do not want to die, I should give up **alcohol**). *Shōsei-mo* means "I also", while written with the same Chinese character as *sei-mo* it can be read *umo* which is one of conjugated forms of *umu* (to give birth) used as a clue word and lack of morphological recognition caused system to incorrectly assume that "straight can be born from alcohol". There was another example for the assertion ($\mathcal{C}_1$) *tōri* (street), ($\mathcal{R}$) "AtLocation", and ($\mathcal{C}_2$) *kuruma* (car). The assertion suggests "street" (*tōri*) can be found at a "car" (*kuruma*), so the concepts ($\mathcal{C}_1$) and ($\mathcal{C}_2$) were naturally negated by the human subjects (cars can be found on the streets, not the opposite). However, the system evaluated the assertion as common, because noun *tōri* was incorrectly matched as a verb which is one of conjugated forms of *tōru* (to pass). This error was caused by the following corpus sentence: *kono-mae, chikaku-wo kuruma-de tōri-mashita.* (recently, I **passed** near by **car**). Above examples show that although it significantly increases the processing time, part of speech information should be added in future.

Another obvious problem is the insufficient corpus size. Even if an assertion represents common sense, it does not always exist in the corpus. We also found problems related not only to concepts ($\mathcal{C}_1$, $\mathcal{C}_2$) but also to relations ($\mathcal{R}$), which co-occur with different objects or subjects in the corpus. For instance, for assertion ($\mathcal{C}_1$) *nattō* (fermented soybeans), ($\mathcal{R}$) "MadeOf", and ($\mathcal{C}_2$) *tōfu* (bean curd), following sentence was retrieved from the corpus: *O-tōfu-mo nattō-mo daisuki-nanode, kondo tsukutte-mimasu* (I'll try to make fermented soybeans and tofu because I love them). Both concepts can be made but there is no relation indicating what is made with what.

6 Discussion and Additional Tests

Considering "All Elements" method, when compared to "MadeOf" and "UsedFor" relations, "AtLocation" reached lower f-score. This is because for "MadeOf" and "UsedFor" assertions we used verbs (and their synonyms) with wide meaning like "*tsukuru*" (to make) and "*tsukau*" (to use), but we did not find their appropriate equivalents for "AtLocation". As presented earlier, we replaced $\mathcal{R}$ with place-indicating particles and added them to concepts: "$\mathcal{C}_2$-*de* $\mathcal{C}_1$" and "$\mathcal{C}_2$-*ni* $\mathcal{C}_1$". However this method did not bring satisfying results (see end of the sections)

For "MadeOf" and "UsedFor" relations f-score is higher for "Concepts Only" than for "All Elements" method due to the higher recall. Taking "UsedFor" relation as example, 53 assertions agreed with human annotators in "All Elements" method, but 12 more correct ones were retrieved when "Concepts Only" method was used. For "MadeOf" relation, our intuition was that retrievals would also be more precise when "All Elements" method is used it was impossible to retrieve correct relations as: ($\mathcal{C}_1$) *shōmei kigu* (lighting equipment), ($\mathcal{R}$) "MadeOf" and ($\mathcal{C}_2$) *garasu* (glass), ($\mathcal{C}_1$) *makura* (pillow), ($\mathcal{R}$) "MadeOf" and ($\mathcal{C}_2$) *menka* (cotton), ($\mathcal{C}_1$) *borushichi* (borscht), ($\mathcal{R}$) "MadeOf" and ($\mathcal{C}_2$) *gyūniku* (beef). However, only in this case precision was lower for $\mathcal{R}$ ($\mathcal{C}_1$, $\mathcal{C}_2$) retrievals (see Table 2).

To improve recall, using only two elements in one sentence is better. However we believe that if the task is to decrease number of assertions for human evaluation, precision is more important. Insufficient corpus and too few appropriate clue words seem to be two main remaining problems. The former is relatively easier to solve by further extension of web-crawling process. On the other hand, the latter is difficult because a concept often depends on context and there is no universal clue word to cover all cases. For example, ($\mathcal{C}_1$) *memo* (note), ($\mathcal{R}$) "UsedFor", and ($\mathcal{C}_2$), *monooboe* (memorizing) did not occur in the corpus together as ($\mathcal{C}_1$, $\mathcal{R}$, $\mathcal{C}_2$), but when we checked ($\mathcal{C}_1$, $\mathcal{C}_2$), the following sentence was found: *Monooboe-no ii hito-hodo memo-wo toru* (The faster learner the more notes he takes). Theoretically we could utilize the verb *toru* (to take) as "UsedFor" clue word for finding other assertions, but this would cause substantial amount of noise because the semantic scope of "to take" is too wide. ($\mathcal{C}_1$) *ōbun* (oven), ($\mathcal{R}$) "UsedFor", and ($\mathcal{C}_2$) *pan-wo yaku* (to bake a bread), did not occur in any sentence. Similarly, in *Haitte sugu, me-no mae-niwa pan-wo yaku obun* (Soon after you enter, in front of you, there will be an oven for baking bread), it would be better to use *yaku* (to bake) instead of *tsukau* (to use).

As shown in the previous section, annotators' evaluation scores differ largely, therefore it is difficult to unambiguously determine if a given evaluation is commonsensical or not. In order to see if the system can be more precise, we repeated evaluation with removed clearly doubtful assertions which were judged from 4 - 7 (see Table 5, 6, 7). Results indicate that with this restriction in "All Elements" method can reach higher precision for all three relations and that "All Elements" achieved higher precision than "Concepts only" method. Consequently, as shown in Table 2, we managed to confirm that the reason why precision of "All Elements" method was lower than in the case of "Concepts Only" method is that annotators' evaluations were highly inconsistent.

To see if we can improve f-score without losing precision, we used separate $\mathcal{C}$-$\mathcal{R}$ pairs for retrieval. For "MadeOf" and "UsedFor" relations, our system counted ($\mathcal{C}_1$, $\mathcal{R}$) and ($\mathcal{C}_2$, $\mathcal{R}$) in the corpus. For ($\mathcal{R}$) "AtLocation", we set *iku* (to go), *kuru* (to come), and *hataraku* (to work) as $\mathcal{R}$ relations, and this method shows capability to improve f-score of the automatic evaluation of assertions. If both expressions ($\mathcal{C}_1$, $\mathcal{R}$) and ($\mathcal{C}_2$, $\mathcal{R}$) occur in the corpus separately, it increases possibility that a given assertion is commonsensical. The results (see Table 8) show that For ($\mathcal{R}$)"MadeOf" and ($\mathcal{R}$) "UsedFor", f-score is higher than for "All Elements" method, but it did not reach the level of "Concepts Only" method. However, for ($\mathcal{R}$) "AtLocation", f-score is relatively higher than other two methods. This shows that whether $\mathcal{C}_2$ stands for place or not plays an important role in evaluating assertions.

7 Conclusion and Future Work

Commonsense knowledge evaluation task is harder than commonsense knowledge acquisition, because for the latter you can acquire relatively high quality as errors look like a small fraction of all retrievals and there is a tendency for ignoring them. However, for evaluation task, more precise judgement is needed

Table 5: Evaluation results for “MadeOf” relations (“All Elements” and “Concepts Only” methods) without doubtful assertions.

MadeOf	Accuracy	Precision	Recall	F-score
All Elements Method $(\mathcal{C}_1, \mathcal{R}, \mathcal{C}_2)$	0.464	**0.870**	0.426	0.571
Concepts Only Method $(\mathcal{C}_1, \mathcal{C}_2)$	0.679	0.837	0.766	0.800

Table 6: Evaluation results for “UsedFor” relations (“All Elements” and “Concepts Only” methods) without doubtful assertions.

UsedFor	Accuracy	Precision	Recall	F-score
All Elements Method $(\mathcal{C}_1, \mathcal{R}, \mathcal{C}_2)$	0.479	**1.00**	0.390	0.561
Concepts Only Method $(\mathcal{C}_1, \mathcal{C}_2)$	0.667	0.903	0.682	0.778

Table 7: Evaluation results for “AtLocation” relations (“All Elements” and “Concepts Only” methods) without doubtful assertions.

UsedFor	Accuracy	Precision	Recall	F-score
All Elements Method $(\mathcal{C}_1, \mathcal{R}, \mathcal{C}_2)$	0.547	**0.765**	0.342	0.473
Concepts Only Method $(\mathcal{C}_1, \mathcal{C}_2)$	0.594	0.630	0.763	0.690

Table 8: Evaluation results for each relations when $(\mathcal{C}_1,\mathcal{R})$ and $(\mathcal{C}_2,\mathcal{R})$ were used.

Relation Method	Accuracy	Precision	Recall	F-score
MadeOf $(\mathcal{C}_1,\mathcal{R})$ and $(\mathcal{C}_2,\mathcal{R})$	0.620	0.786	0.705	0.743
UsedFor $(\mathcal{C}_1,\mathcal{R})$ and $(\mathcal{C}_2,\mathcal{R})$	0.550	0.872	0.513	0.646
AtLocation $(\mathcal{C}_2,\mathcal{R})$	0.590	0.619	0.696	0.650

to deal not only with those errors from acquisition systems but also with often wrong input from human annotators.

In this paper we present a new text-mining approach for automatic commonsense knowledge evaluation. “All Elements” method using both concepts and their relation achieved precision of over 70% on average for the three following ConceptNet relations: “MadeOf” (78.0%), “UsedFor” (100.0%) and “AtLocation” (61.5%). We described how different concepts and relation combinations can be utilized and showed their strengths and weaknesses. From the error analysis we revealed main problems which are database contributors originality, the insufficient corpus size, discrepancies in evaluators’ opinions, and setting proper clue words. Especially the first problem shows that it is often hard to evaluate concepts stored in their current form. To solve it, instead of using a concept as it is, its more frequently used synonymic concepts should be utilized. For example, in the case of assertion $(\mathcal{C}_1)$ *shōmei kigu* (lighting equipment), $(\mathcal{R})$ “MadeOf”, and $(\mathcal{C}_2)$ *garasu* (glass), our system could search for “lamp” instead of the “lighting equipment” (there were 11 hits instead of 0 when we tried this for “All Elements” method). In near future, we plan to increase the number of annotators, because commonsense knowledge differs depending on subjects and their particular experiences. We will also experiment with different clue words for higher recall without losing precision.

Our methods are also planned to be utilized in commonsense knowledge acquisition system as its self-evaluation module. We are also going to test our idea in different languages used in ConceptNet.

References

Francis Bond, Hitoshi Isahara, Sanae Fujita, Kiyotaka Uchimoto, Takayuki Kuribayashi, and Kyoko Kanzaki. 2009. Enhancing the Japanese WordNet. In *Proceedings of the 7th workshop on Asian language resources*, pages 1–8. Association for Computational Linguistics.

Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R Hruschka Jr, and Tom M Mitchell. 2010. Toward and architecture for Never-Ending Language Learning. In *AAAI*, volume 5, page 3.

Hyemin Chung. 2006. *GlobalMind- -Bridging the Gap Between Different Cultures and Languages with Commonsense Computing*. Ph.D. thesis, Massachusetts Institute of Technology.

Jonathan Gordon, Benjamin Van Durme, and Lenhart K Schubert. 2010. Evaluation of commonsense knowledge with Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 159–162. Association for Computational Linguistics.

Douglas B Lenat. 1995. Cyc: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11):33–38.

Kazuhiro Nakahara and Shigeo Yamada. 2011. Development and evaluation of a Web-based game for commonsense knowledge acquisition in Japan (in Japanese). *Unisys Technology Review*, (107):295–305.

Stellan Ohlsson, Robert H Sloan, György Turán, Daniel Uber, and Aaron Urasky. 2012. An approach to evaluate AI commonsense reasoning systems. In *FLAIRS Conference*, pages 371–374.

Michal Ptaszynski, Pawel Dybala, Rafal Rzepka, Kenji Araki, and Yoshio Momouchi. 2012. Annotating affective information on 5.5 billion word corpus of Japanese blogs. In *Proceedings of The 18th Annual Meeting of The Association for Natural Language Processing (NLP*.

Rafal Rzepka, Koichi Muramoto, and Kenji Araki. 2011. Generality evaluation of automatically generated knowledge for the Japanese ConceptNet. In *Australasian Joint Conference on Artificial Intelligence*, pages 648–657. Springer.

Lenhart Schubert. 2002. Can we derive general world knowledge from texts? In *Proceedings of the second international conference on Human Language Technology Research*, pages 94–97. Morgan Kaufmann Publishers Inc.

Push Singh, Thomas Lin, Erik T Mueller, Grace Lim, Travell Perkins, and Wan Li Zhu. 2002. Open Mind Common Sense: Knowledge acquisition from the general public. In *OTM Confederated International Conferences "On the Move to Meaningful Internet Systems"*, pages 1223–1237. Springer.

Robert Speer and Catherine Havasi. 2012. Representing General Relational Knowledge in ConceptNet 5. In *LREC*, pages 3679–3686.

Niket Tandon, Gerard de Melo, Fabian Suchanek, and Gerhard Weikum. 2014. WebChild: harvesting and organizing commonsense knowledge from the Web. In *Proceedings of the 7th ACM international conference on Web search and data mining*, pages 523–532. ACM.