Automatic selection of predicates for common sense knowledge expression

Ai MAKABI
Department of Electrical Engineering
Nagaoka University of Technology
Nagaoka City, Niigata, Japan
makabi@jnlp.org

Hiroshi MATSUMOTO
Department of Electrical Engineering
Nagaoka University of Technology
Nagaoka City, Niigata, Japan
matsumoto@jnlp.org

Kazuhide YAMAMOTO
Department of Electrical Engineering
Nagaoka University of Technology
Nagaoka City, Niigata, Japan
yamamoto@jnlp.org

***Abstract*—We are constructing common sense knowledge base that includes common sense knowledge such as a dog walks on four legs. We use Web text as a source information for construction. One may expect to obtain so many common senses from Web text, but we also obtain so many facts together that are not common senses, hence it is required to select common senses from those. In this paper we propose a method for statistically selecting common senses of nouns utilizing the unique number of co-occurred predicates. We illustrate that, using only frequency of predicates, the method can acquire common sense knowledge properly.**

***Keywords*-common sense knowledge base; similarity of noun-predicates pairs; knowledge extraction; common sense representation;**

I. Introduction

Recently, the deep semantic analysis, expressing the meaning of words using the common sense knowledge, are actively researched. The "*common sense knowledge*" we are referring to here are the collection of facts and information of which people are expected to know. For instance, sentences of "A dog walks on four legs" and "The display can show an image" are examples of the common sense knowledge. We need not only grammatical knowledge, but also a large amount of common sense knowledge in order to develop the intelligent computer. Therefore, many researchers focus on the research which build the common sense knowledge base, and provide them to any natural language processing tasks in the accessible representation.

In this study, we aim at a construction of easy-to-use Japanese common sense knowledge base for semantic analysis in natural language processing. First of all, we define that predicates(verbs, adjectives, verbal nouns) which co-occur with a noun are the common sense knowledge of the noun. The term *verbal noun*, or what we call *sahen noun*, is subgroup of noun which is also used as verb when followed by a suffix "suru". Based on the definition, we propose a automatic extraction method of common sense knowledge from a large-scale Web text. For example, when predicates of "ほえる (to bark)" and "散歩 (to take out for a walk)" are co-occurred with noun of "犬 (dog)" on Web text, both of them are common sense knowledge of "犬 (dog)". We can analyze each noun at the predicate level for using the common sense knowledge base, such as analyzing relationships among nouns by calculating the similarity of the predicate set.

But of course, all of the predicates which co-occurs with noun are not appropriate as common sense knowledge. Whether a predicate is a correct common sense or not, it is different for each noun. Hence, in this paper, we describe how to select the appropriate predicates as common sense knowledge for each noun in order to construct the common sense knowledge base.

II. Related work

In artificial intelligence research, the common sense knowledge base is also called "upper ontology (or top-level ontology)". The *upper ontology* is an ontology which contains many general concepts, and most of other ontologies are accessibly ranked under the upper ontology. For example, SUMO(Suggested Upper Merged Ontology)[1] and Cyc[2] are well known as representative of upper ontologies. There are much research based on each upper ontology in natural language processing tasks. Ahrens et al.[3] integrated the Conceptual Mapping Model with SUMO in order to demonstrate for using conceptual metaphor analysis. Moreover, in other works(Niles et al.[4], Jan et al.[5], Reed et al.[6]), they mapped synsets of existing linguistic resources(WordNet[1], FrameNet[2], and so on) to SUMO concepts or Cyc concepts, and attempt to construct the generic knowledge base. The study using upper ontologies is exploitable rigorously-defined common sense knowledge, but on the other hand, knowledge representation defined by upper ontology can not yet fully be corresponded with actual expressions.

Compared to this, the ConceptNet[3][7], semantic network provided by Open Mind Common Sense Project(OMCSP), defined the common sense knowledge of each concept as sentences or words adding some relations(e.g. *IsA*, *CapableOf*, *RelatedTo*). It is better suited to a natural language task than the upper ontology because the common sense knowledge is defined as natural languages. But in Japanese,

[1]http://wordnet.princeton.edu/
[2]https://framenet.icsi.berkeley.edu/fndrupal/
[3]http://conceptnet5.media.mit.edu/

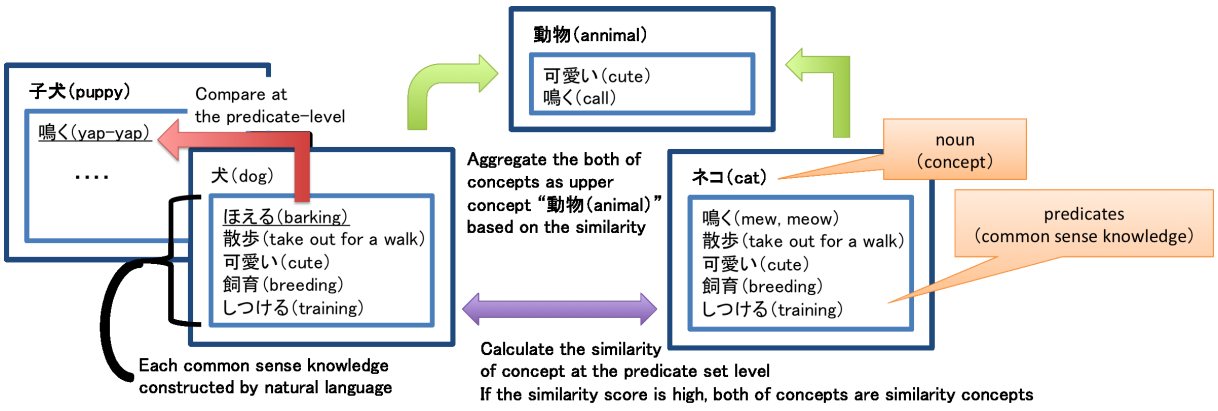


Figure 1. Overview of the common sense knowledge base which would be a goal

most of concepts are collected by hand, so coverage of common sense knowledge is exceptionally-low. Actually, the number of common sense knowledge stored in the ConceptNet are 1,035,681 sentences in English, and 14,546 sentences in Japanese. Rzepka et al. tried to automatically generate the common sense knowledge based on the Japanese ConceptNet, but they have not obtained the adequate result yet[8].

As a similar system, Chklovski[9] construct the large amount of knowledge base using the cumulative analogy which is a class of analogy-based reasoning algorithms. But the system are also constructed by hand, so we should pay the high construction costs.

Hence in this study, we express common sense knowledge in natural language in order to build the common sense knowledge base which is accessible for natural language processing, and aim at automatic construction of the knowledge base.

III. Automatic Selection of Predicates

In this study, we define the predicates characterizing the noun as common sense knowledge, and make the following hypothesis as specific property of them.

(1) The predicate a is the common sense knowledge of the noun n when the pair of a and n are frequently co-occurred in sentences.
(2) The predicate n which co-occurs with any nouns is not the appropriate common sense knowledge because the noun is characterized by the set of common sense knowledge.
(3) Whether the predicate a is a correct common sense or not, it depends on the number of unique nouns which co-occurred with a. For the predicate which co-occurs with many nouns, if the predicate has nouns which co-occur with only a few other predicates, the predicate is considered as common sense knowledge. [4]

Based on the hypothesis, we remove incorrect predicates from the co-occurrence predicates with noun, and should select appropriate predicates.

As analysis objects, we use the pairs of noun and predicate which are co-occurred in the 7-gram data. The *7-gram data* is a part of "Web Japanese N-gram[KK]" which includes the word N-grams parsed by morphological analyzer *MeCab*[mec]. Each N-gram appeared more than 20 times from 20 billion sentences in Web text. The sum of the unique 7-grams including in the 7-gram data is 570,204,252.

First of all, we analyze constituent morphemes using MeCab, and extract the co-occurring pairs of noun and predicate. Then, we automatically regularize the spelling variations by using a spelling unification dictionary[spe]. that includes 28,810 words which have some spelling variations. As a result, we could extract 605,363,630 noun-predicates pairs (The unique number of pairs is 29,434,191, consisting of 655,038 nouns and 26,455 predicates).

Next, we sort the nouns by number of co-occurring predicates based on the hypothesis (1), and investigate the emergence distribution of predicates in the top N nouns. Figure 2 shows the emergence distribution of predicates in the top 1,000 nouns. The horizontal axis indicates the number of unique nouns co-occurring with predicate. For example, the number of unique nouns is 500 when a predicate "走る (run)" co-occurred with 500 unique nouns in the top N=1,000 nouns. The vertical axis means the number of unique predicates with logarithmic expression. For example, the number of unique predicates is 10 when there are 10 predicates whose number of unique nouns is 500.

As the emergence distribution, we realize that the number of unique predicates dramatically increases when a number of unique nouns is extremely large or few. Since the pred-

[4]e.g. the predicate "run" could not characterize the noun "person" which frequently co-occurred with other predicates, but it could characterize the "runner" which occurred with only a few other predicates.

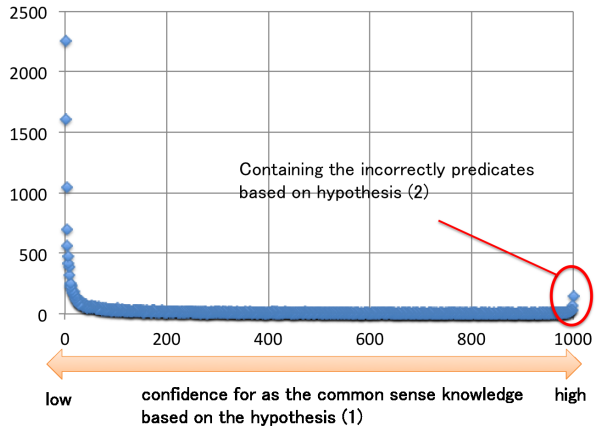


Figure 2. Emergence distribution of top the 1,000 predicates co-occurring with noun (horizontal axis: a number of unique nouns co-occurring with predicates, vertical axis: a number of unique predicates)

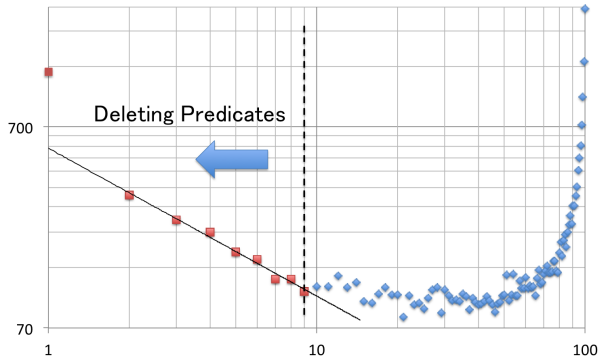


Figure 3. The emergence distribution of predicates and the power approximated curve of figure 4 (1) (horizontal axis: a number of unique nouns co-occurring with predicates (logarithm), vertical axis: the number of the number of unique predicates (logarithm))

icates with high co-occurence frequency with the noun are the common sense knowledge with a high probability under the hypothesis (1), thus the more the predicate is distributed on right side, the higher probabilty of the common sense knowledge.

Moreover, under the hypothesis (2), we see the predicates which co-occurred with any nouns as incorrect common sense knowledge, then they are deleted. In this study, we focus on the point of which the number of predicates is sharply increased when the number of unique nouns is large, and the predicates, which fall under a certain scope, are as the deleting predicates. We decide the scope of deleting predicates for a list of points on a power approximated curve (The part shown by a red circle in Figure 2). Figure 3 shows the emergence distribution of predicates and the power approximated curve in the top 1000 nouns. It is a double logarithmic plot, so the approximated curve which means a dramatically increasing area shows as straight line. We calculate distances of line and points, and decide whether a scope of point includes appropriate predicates or not. The red points on the figure shows the scope of deleting predicates.

Finally, we examine the hypothesis (3). It means that the noun which co-occurs with many predicates can not be characterized by generic predicates, hence, the number of deleting predicates for them increase. In fact, when we sort the nouns by the number of co-occurring predicates (the more nouns ranked upper, they have more deleting predicates). Considering the above, we investigate the emergence distribution of top the N predicates co-occurring with noun.

Figure 4 shows the emergence distributions of predicates in the top N nouns , where N=100, 1,000, and 10,000 respectively. Thus, high-ranked nouns will have more deleting predicates when nouns are sorted in the order of predicate-cooccurence. Hence, this matches with the hypothesis (3).

Based on the result, we make an investigation of changes in deleting predicate while increasing N until there is no more increase. Figure 5 shows how the number of deleting predicates change from N equals 100 to 4,500. The number of deleting predicates decreased in a staircase pattern, and there are singular points in N=700, 1,100, 1,600, 2,500 or 3,600. Based on this result, we decide the number of deleting predicates for each noun under the hypothesis (3).

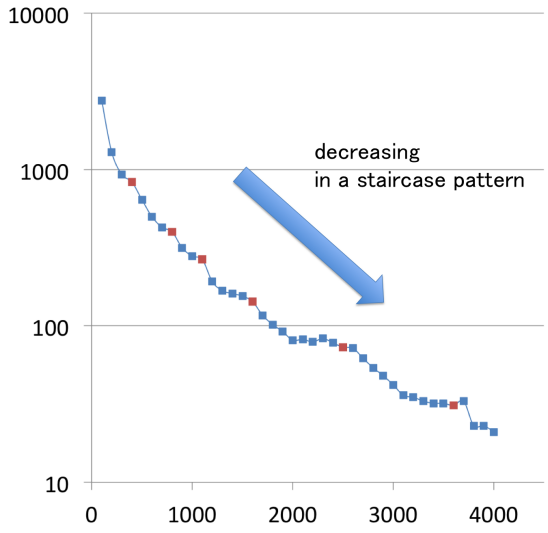
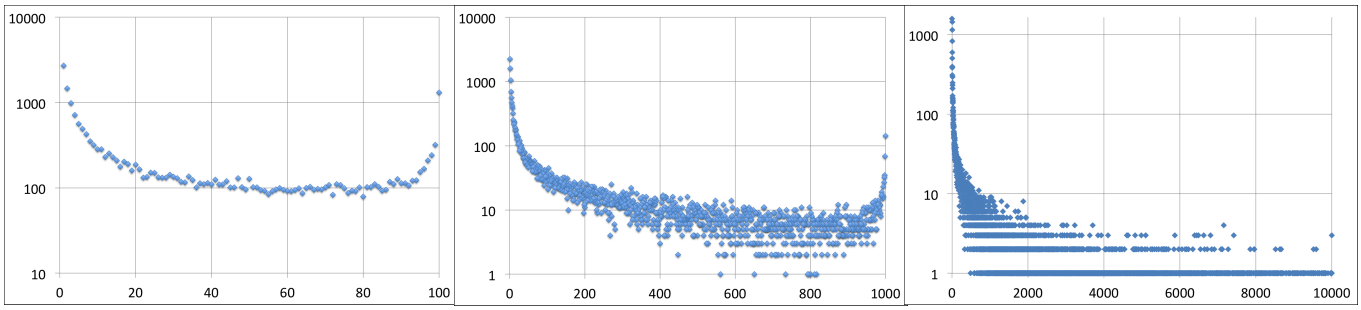


Figure 5. The number of deleting predicates changes from N equals 100 to 4,500 (horizontal axis: The top N nouns co-occurring with many predicates, vertical axis: The number of deleting predicates(logarithm))

Table I shows the number of deleting predicates for each noun. For example, 227 predicates are deleting target for nouns which N is 1,000.



(1) N=100 (2) N=1,000 (3) N=10,000

Figure 4. Emergence distribution of top the N predicates co-occurring with noun (horizontal axis: a number of unique nouns co-occurring with predicates, vertical axis: a number of unique predicates (logarithm))

Table I
THE NUMBER OF DELETING PREDICATES FOR EACH NOUN (N=THE UNIQUE NUMBER OF CO-OCCURRED PREDICATES)

Scope of the nouns	Deletion
N≤700	427
700<N≤1,100	267
1,100<N≤1,600	143
1,600<N≤2,500	73
others	33

However, the 33 predicates, which get deleted when N=3,600, can be used to nearly all nouns, so we consider that they are not common sense knowledge, and delete from all nouns as incorrectly predicates. Figure 6 shows a list of the deleting predicates for all nouns.

わかる (understand), もつ (have), みる (see, look), なる (become), ない (nothing), とる (take, adopt, prefer), できる (can), つく (get, prick, arrive, strike), しる (know), くる (come), おもう (think), おおい (many), いる (be, need, shoot), いう (say), ある (be), 良い (good), 入る (enter, join), でる (leave, go out, attend), つくる (make), つかう (use), きく (hear, ask, be effective), かく (write, scratch), おこなう (do), 紹介 (introduce), よい (good), ゆく (go), たつ (stand, build, pass), たかい (high, expensive), おる (be, fold), いい (good), 関係 (to relate), やる (do), かける (build, hang, run, lack)

Figure 6. The deleting predicates for all noun

We use the selected predicates as common sense knowledge, and add them to each noun. In particular, we calculate the weighted scores for predicates co-occurring with noun using Harman normalized frequency. A predicate is correct common sense knowledge for a noun when the predicate score is high. The equation of Harman normalized frequency is as follows (n: noun, a: predicate, $n_{a,n}$: appearance frequency of predicate a with noun n).

$$TF(a,n) = \frac{\log_2(n_{a,n}+1)}{\log_2(\sum_k n_{k,n})} \quad (1)$$

IV. EVALUATION

A. Evaluation method

We compare the adding common sense knowledge to following baselines.

(1) Do not delete the any predicates, just use the weighted predicates by Harman normalized frequency (baseline1).
(2) Do not delete the any predicates, just use the weighted predicates by the equation 2 based on TF-IDF (baseline2).
(3) Remove the 427 deleting predicates in N≤700 shown by Table I, and use the weighted predicates by Harman normalized frequency (baseline3).

We compare our method with baseline1 and baseline2 to verify the effect of deleting predicates which frequently co-occurred with unique nouns. Moreover, for comparing approach method to baseline3, we check the effect of shifting numbers of deleting predicates for each noun.

The following equation shows the weighted score for a predicate a which co-occurred with a noun n used by baseline2 ($|N|$: the sum of nouns, $|N_a|$: the sum of nouns which co-occurred with a predicate a).

$$wgt(a,n) = TF(a,n) \times \left\{ \log_2 \frac{|N|}{|N_a|} + 1 \right\} \quad (2)$$

B. Evaluation Result

We take three different nouns for evaluation, and show their assigned predicates which ranked in the top 10 as follows (Table II).

The proposed method can assign appropriate predicate to a noun as the common sense knowledge for the most nouns. On the other hand, in baseline1 and baseline2, a predicate which frequently co-occurred with any nouns are ranked much higher than proposed method. For example of a noun "犬 (dog)", predicates of "なる (become)", "いる (be)" and "一緒 (be together)" which co-occurred with any nouns are appeared in higher rank in baseline1 and baseline2. However, in approach method, incorrect predicates are deleted

Table II
THE DIFFERENCE OF THE TOP 10 PREDICATES ADDING TO THE NOUNS

noun: いぬ (dog)			
Baseline1	Baseline2	Baseline3	Proposed
かう (buy, have)	かう (buy, have)	喰わない (don't eat)	散歩 (to take out for a walk)
なる (become)	一緒 (be together)	飼わない (don't breed)	しつける (bring up)
いる (be)	生活 (to live)	かみころす (bite to death)	病気 (be sick)
ある (be, stand)	販売 (to sale)	吠えない (don't bark)	つれる (take someone to tow)
生活 (to live)	たのしい (fun)	薬殺 (to give a lethal injection)	くらす (live)
みる (see, find)	やすい (cheap, easy)	繋留 (to tether)	訓練 (to train)
ない (be none)	わかる (understand)	訓練 (to train)	ほえる (bark)
いう (say)	登録 (to register)	やせこける (get all thin)	かわいい (cute)
一緒 (be together)	大きい (big)	かまない (don't bite)	介護 (to care for)
できる (be able to)	かんがえる (think)	代参 (to participate instead)	飼育 (to breed)
noun: 小学校 (elementary school)			
Baseline1	Baseline2	Baseline3	Proposed
入学 (to enroll in school)	就学 (to attend school)	離任 (to leave a school)	入学 (to enroll in school)
教育 (to educate)	入学 (to enroll in school)	訓導 (to teach, to lead)	教育 (to educate)
ある (be)	付属 (to attach)	めざめない (don't arise from sleep)	卒業 (to finish school)
なる (become)	参観 (to make a classroom visit)	さかしい (intelligent)	授業 (to give lessons)
卒業 (to finish school)	給食 (to eat school lunch)	加減乗除 (to be four operations)	受験 (to take an exam)
授業 (to give lessons)	受験 (to take an exam)	そばだつ (rise, tower)	かよう (attend)
受験 (to take an exam)	授業 (to give lessons)	歌わす (get to sing)	学習 (to learn)
かよう (attend)	担任 (to take charge of)	やり直さない (don't start again)	指導 (to coach)
学習 (to learn)	卒業 (to finish school)	のびゆく (would glow)	依頼 (to ask, to request)
指導 (to coach)	かよう (attend)	実験 (to do a experiment)	就学 (to attend school)
noun: 赤ちゃん (baby)			
Baseline1	Baseline2	Baseline3	Proposed
やすい (cheap, be easy)	うまれる (be born)	すわる+ない (don't sit)	うまれる (be born)
できる (be able to, be born)	やさしい (gentle)	さずかる (be blessed with)	ほしい (want)
たのしい (fun)	たんじょう (be born)	ぐずる (be peevish)	できる+ない (can't, be not born)
きたる (come)	ほしい (want)	沐浴 (to take a ritual bath)	大きい (big)
あわせる (join, adopt)	一緒 (be together)	ほ乳 (to suck)	ねる (sleep)
いる (be)	大きい (big)	出産 (be born)	あう (met)
入れる (put in)	のむ (drink)	発育 (glowing)	えらぶ (select)
かんがえる (think)	あう (met)	うまれる (be born)	できる (be able to, be born)
みる (look, see)	やすい (cheap, be easy)	授乳 (nursing)	たのしい (fun)
かえる (change)	記念 (to commemorate)	うむ (have a baby)	経験 (to be experience)

and appropriate predicates such as "散歩 (to take out for a walk)" and "しつける (bring up)" are appeared in higher rank. The proposed method acquires better performance than baseline1 and baseline2. As a result, we confirm the effect of deleting predicates which co-occur with many nouns. We can do the meticulous comparison at the predicate level, it is much better than previous studies which automatically construct a noun thesaurus using verbs [10][11].

Moreover, baseline3 deleted some appropriate predicates that consequently remains some incorrect predicates in the top ranking. For example, appropriate such as "入学 (to enroll in school)" and "学習 (to learn)" are deleted, and incorrect predicates such as "めざめない (don't arise from sleep)" and "さかしい (intelligent)" are appeared in higher rank. For this reason, the proposed method, changing the number of deleting predicates for each noun, is absolutely proper.

C. Error analysis

The table III shows examples of failure in predicate-assignment. Although a predicate co-occurs with a noun many times, there are sometimes unrelated pairs because we don't check the dependency relation between them. It is the primary cause of failure. For this solution, we should use only the predicates which depend on the target nouns as candidates of common sense knowledge.

It is a rare case that we can not assign nouns, which can also be used as suffix (such as "月 (month)"), to appropriate predicates (e.g. 6 月に入籍する (We get married in June), 月ごとに決済する (We make a charge for each month)). In point of "月 (month)", it also means a moon, hence we need to determine the senses. This problem can be eased by utilizing the relation of another co-occurred nouns (e.g. If the "月" is co-occurred with a noun "太陽 (sun)", it may

Table III
THE FAILURE EXAMPLES OF ADDING PREDICATE

名詞：月 (moon, month)	名詞：理由 (reason)	名詞：原因 (cause)
必着 (must arrive)	やむをえない (unavoidable)	救命 (to save someone's life)
決算 (to settle an account)	返品 (to return articles)	老化 (to age)
施行 (to enforce)	稼げない (can't earn)	つきとめる (pinpoint, discover)
公布 (to promulgate)	拒絶 (to refuse)	故障 (to break down)
利上げ (to raise the rate of interest)	解雇 (to discharge)	ひきおこす (cause, give)
ずれこむ (drag on into)	削除 (deleting)	病気 (to contract an illness)
入籍 (to enter in)	志望 (to want to)	食中毒 (to have food poisoning)
ぞくする (belong to)	却下 (to reject)	出火 (to outbreak of fire)
連結 (to connect)	ことわる (decline)	肥満 (to fat)
着工 (to start)	上告 (to enter an appeal)	くすむ (be darkish)

mean the moon).

Furthermore, nouns of "理由 (reason)" and "原因 (cause)" are inadequate as adding target of common sense knowledge because they are used for defining the relation of nouns. In the future, we should discuss how we limit the nouns of adding target.

V. CONCLUSION

In this paper, we described the selection method of appropriate predicate as common sense knowledge for constructing the common sense knowledge base. We sorted the nouns by number of co-occurring predicates, and investigated the relation of appearance frequency of the top N nouns and predicates. As a result, we decide some predicates as deleting target for the nouns in the specified scope (N=700, 1,000, 1,600, 2,500, and 3,600).

We evaluated sets of common sense knowledge which are assigned to each noun, and realized that our proposed method can add appropriate predicates to the noun compared with three baselines. As a result, we demonstrated assumed characteristics of common sense knowledge in our study. That is to say, a predicate which co-occurs with any nouns is not the appropriate common sense knowledge, and whether a predicate is a correct common sense or not, it depends on the noun which co-occurred with the predicate.

TOOLS AND LANGUAGE RESOURCES

[KK] Taku Kudo and Hideto Kazawa. Web japanese n-gram version 1. published by Gengo Shigen Kyokai.

[mec] Mecab 0.993. http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html.

[spe] Spelling unification dictionary. http://www2.ninjal.ac.jp/lrc/index.php?%A1%D8%C9BD%B5%AD%C5%FD%B9%E7%BC%AD%BD%F1%A1%D9.

REFERENCE

[1] I. Niles and A. Pease, "Towards a standard upper ontology," in *Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001)*, 2001, pp. 17–19.

[2] D. B. Lanat, "Cyc: a large-scale investment in knowledge infrastructure," in *Commun. ACM*, vol. 11, 1995, pp. 33–38.

[3] K. Ahrens, S. Chung, and C. Huang, "Conceptual metaphors: Ontology-based representation and corpora driven mapping principles," in *Proceedings of the ACL 2003 workshop on Lexicon and figurative language*, vol. 14. Association for Computational Linguistics, 2003, pp. 36–42.

[4] I. Niles and A. Pease, "Linking lexicons and ontologies: Mapping wordnet to the suggested upper merged ontology," in *Proceedings of the IEEE International Conference on Information and Knowledge Engineering*, 2003, pp. 412–416.

[5] H. Hennett and C. Fellbaum, "Linking framenet to the suggested upper merged ontology," in *Formal Ontology in Information Systems: Proceedings of the Fourth International Conference (Fois 2006)*, vol. 150. Ios PressInc, 2006, p. 289.

[6] S. L. Reed, D. B. Lenat *et al.*, "Mapping ontologies into cyc," in *AAAI 2002 Conference Workshop on Ontologies For The Semantic Web*, 2002, pp. 1–6.

[7] C. Havasi, R. Speer, and J. Alonso, "Conceptnet 3: a flexible, multilingual semantic network for common sense knowledge," in *In Recent Advances in NLP*, 2007.

[8] R. Rzepka, K. Muramoto, and K. Araki, "Generality evaluation of automatically generated knowledge for the japanese conceptnet," pp. 648–657, 2012.

[9] T. Chklovski, "Learner: a system for acquiring commonsense knowledge by analogy," in *Proceedings of the 2nd international conference on Knowledge capture*. ACM, 2003, pp. 4–12.

[10] D. Hindle, "Noun classification from predicate-argument structures," in *Proceedings of the 28th annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 1990, pp. 268–275.

[11] M. Hagiwara, Y. Ogawa, and K. Toyama, "A comparative study on effective context selection for distributional similarity," in *Journal of Natural Language Processing*, vol. 5, no. 5, 2008, pp. 119–150.