Fusion of Multiple Semantic Networks and Human Association

Hitoshi Isahara
Toyohashi University of Technology
Toyohashi, Aichi, Japan
isahara@tut.jp

Kyoko Kanzaki
Toyohashi University of Technology
Toyohashi, Aichi, Japan
kanzaki@imc.tut.ac.jp

Eiko Yamamoto
Gifu Shotoku Gakuen University
Gifu, Japan
eiko@gifu.shotoku.ac.jp

Takayuki Kuribayashi
Toyohashi University of Technology
Toyohashi, Aichi, Japan.
kuribayashi@lang.cs.tut.ac.jp

Michinaga Otsuka
Toyohashi University of Technology
Toyohashi, Aichi, Japan
michinaga@lang.cs.tut.ac.jp

Abstract

We are trying to construct a conceptual system that accurately represents human thoughts by fusing of semantic networks. As semantic networks to fuse, we use the Japanese Wordnet which is a thesaurus made manually based on linguistic intuition and the knowledge acquired automatically from the actual text stored in the huge corpus. Such knowledge are represented as mutual relations of the concepts of words. In order to acquire such relations, we focus on the case relations in sentences and calculate inclusive relations of co-occurrence by using Complementary Similarity Measure. As an application and verification of the conceptual system created, we try to simulate human associations by using the conceptual system. As an experimental result, we found the obvious difference in generated association links between using the semantic network of Japanese Wordnet and using the fused semantic networks with Japanese Wordnet and the acquired mutual relations.

1 Introduction

In systems that support human creativity and search, a dictionary data similar to human perception is required. Human do not think in only classification knowledge. It is insufficient for the systems which support human cognitive processes to utilize only those existing language resources such as thesauri that summarize word senses and conceptual relationships of words. Because humans express their thoughts with words, it is valid to acquire knowledge from their actual utterances and contexts reflecting their thoughts.

In this study, we are trying to construct a conceptual system that accurately represents human thoughts by fusing of semantic networks. As semantic networks to fuse, we use the following two kinds of knowledge structure. As the first one, we used the Japanese Wordnet (Isahara et al, 2008) which is a thesaurus made manually based on linguistic intuition. As the second one, we use the knowledge acquired automatically from the actual text stored in the huge corpus. Such knowledge are represented as mutual relations of the concepts of words. In order to acquire such relations, we focus on the case relations in sentences such as "case and statement," "verb and object" and "subject and verb," and calculate inclusive relations of co-occurrence by using Complementary Similarity Measure (CSM) (Hagita and Sawaki, 1995; Yamamoto et al., 2005).

As an application and verification of the conceptual system created, we try to simulate human associations by using the conceptual system. Concretely, we first conduct an experiment on the association with a stimulus word, and create the association network based on the experimental result by using our conceptual system. Then, we visualize the structure of the created association networks and analyze them as networks. As an experimental result, we found the obvious difference in generated association links between using the semantic network of Japanese Wordnet and using the fused semantic networks with Japanese Wordnet and the acquired mutual relations.

2 Experimental Data

To realize our aim described above, we create new knowledge structure by combining the Wordnet which is manual made thesaurus with taxonomical information and the mutual relations between words which is extracted from actual text in a huge web corpus. In this section, we explain data for our experiment.

2.1 Japanese Wordnet

As for Wordnet, we use Japanese Wordnet version 1.1, whose specifications are shown in Table 1.

Number of words	93,834
Number of senses	158,058
Number of synlinks	283,600
Number of synset	57,238
Number of gloss	135,692
Number of example sentence	48,276

Table 1. Specifications of Japanese Wordnet

2.2 CSM data

In this study, we utilize the knowledge based on human utterances to construct a semantic network as a representation of human thought. We use Complementary Similarity Measure (CSM) to acquire such knowledge from the actual text.

CSM is an asymmetry and noise-resistant measure. Values obtained by CSM indicate relations between words, such as Hypernym-Hyponym. We named data obtained in these process "CSM data." Comparing the Japanese Wordnet and the CSM data, we found a lot of words and word relations that retrieved from web corpus but that have not been stored in the Wordnet. Therefore, we constructed new conceptual system based on the Japanese Wordnet that enriched by conceptual relationships with word pairs in CSM data.

2.3 Experimental data based on case relation

In our experiment, we use web corpus with 500 million Japanese sentences (Kawahara and Kurohashi, 2006). We analyze syntactically 500 million sentences and extract pairs of words having co-occurrence relations in an actual sentence by focusing on case relation, namely modified/modifier relationship. Then, we calculate CSM value for each pairs, after we reduce some noises in the extracted pairs by setting a threshold value.

To estimate inclusive relations between words, we applied the method based on the CSM, which estimates inclusive relations between two vectors (Yamamoto et al., 2011). By using an appearance pattern as a feature vector for each word in treating linguistic resource such as a corpus or document collection, we have reported being able to determine a relation between two words according to the inclusive relation estimated by the CSM value.

The Japanese language has case-marking particles that indicate the semantic relation between two words in a dependency relation. Then, using the syntactic analysis result of the web corpus, we collected words in case (dependency) relations.

We considered the meaning of some case relations as follows.

Subject and Verb (SV)

The set of verbs that occur with certain kinds of nouns as their subject represents the behavior of the noun. To extract this relation, we use case-marking particle <ga>.

For example, if "dog" occurs with "eat" and "bear," and "animal" occurs with "fly", "eat", and "bear", then we considers that "dog is an animal" since the behavior of "dog" is a subset of (or included in) the behavior of "animal."

Verb and Object (VO)

The set of verbs that occur with certain kinds of nouns as their object represents "how to treat." To extract this relation, we use case-marking particle <wo>.

For example, as "criminal" often appears as an object of the verb "catch," we can imagine that a criminal is a person who tend to be caught.

Noun and Sentence (NS)

For each noun N in a sentence S, we can regard that N co-occurs with S. In other words, nouns appearing in same sentence have a relationship each other. They tend to be together in one specific scene in a real world. In our experiment, we extracted such relations by gathering nouns in a sentence with case-marking particles <ga>, <no>, <wo>, and <de>.

	NS	SV	VO	SO-S	SO-O
Number of extracted words	4,676,041	1,449,150	1,503,255	395,734	346,531
Threshold	10	2	3	2	2
Number after elimination	246,717	176,511	114,336	31,531	32,703
Number of links with positive CSM value	19,279,434	1,908,489,076	718,477,958	27,801,885	46,351,392

SO-S means nouns in subject position classified by similarities of nouns in object position in a sentence. SO-O is vice versa.

Table 2. Statistics of extracted words and relations

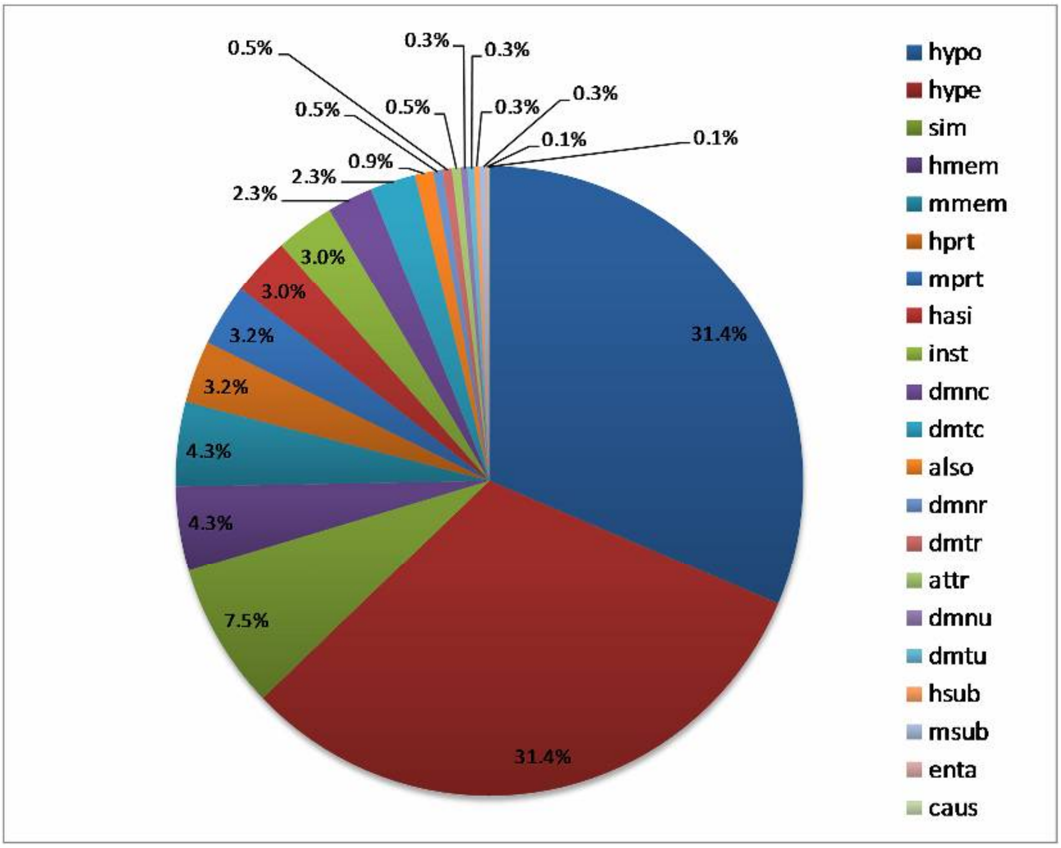

Figure 1. Breakdown of links in Japanese Wordnet

Subject and Object (SO)

This co-occurrence relation is the combination of a subject and an object for same verb.

For example, in the sentence "a human eats a bread", "human" and "bread" are extracted as a combination.

As described above, we calculate the similarity between two words based on the word co-occurrence by using CSM. To do this, we represent by a binary vector experimental data which extracted from corpora based on the case relation; the vector corresponds to the appearance pattern of a noun. We apply the CSM for the calculation (Yamamoto et al. 2005). Parameters for calculating the CSM-value correspond to the number of dimensions in each situation.

Table 2 shows the number of extracted words, the number of extracted words after the elimination by a threshold, its threshold (number of occurrence) and combination of words which has positive CSM value, for each case relation.

In this paper, we use NS data, because it could extract enough number of data with variety of CSM values.

3 Comparison between Japanese Wordnet and CSM data

In this section, we compare Japanese Wordnet with relation between concepts and CSM data which consists of links between two words with its CSM value.

As shown in Table 1, there are 286,300 synlink entries in Japanese Wordnet. Among them, 178,178 entries (63%) are taxonomical links such as hyponym and hypernym. Figure 1 shows the statistics of links in Japanese Wordnet.

As shown in Table 2, we extracted 19,279,434 links which have positive CSM value from our experiments. We chose top 5% of these links by setting a CSM value as a threshold (926,653 links). We use this extracted and eliminated CSM data, i.e. word links with CSM value, for comparison with Japanese Wordnet. The result of comparison are shown in Table 3.

In Table 3, "Number of data" means that the number of links (or relations) extracted automatically by our system. "Percentage for all CSM data" is the percentage of the data among all extracted and eliminated CSM data.

"No wordid" means that one or two words related to this link of CSM data are not stored in Japanese Wordnet. As shown in Table 3, about 63% of data contain words which are not stored in Japanese Wordnet. It shows that CSM data which was extracted automatically from huge corpus are useful to improve the coverage of vocabulary which appears in the real text.

"No synlink" means there is no relation between two synsets in Japanese Wordnet, which correspond to two words in each CSM data. 37% of CSM data are categorized into this class. This category means we can add new relations (links) into Japanese Wordnet based on the cooccurrency between words in the huge corpus.

"Same synset" means that two words in CSM data are treated as synonyms in the Japanese Wordnet. "Hypernym," "hyponym" and others, which are not shown in Table 2, means that two words are already stored in Japanese wordnet properly. We found 1,415 such relations by this experiments, i.e. Hypernym (578), Hyponym (475) and others (362).

Relation	Number of data	Percentage for all CSM data
No wordid	582555	62.8666
No synlink	341868	36.8928
Same synset	815	0.08795
Hypernym	578	0.06238
Hyponym	475	0.05126

Table 3. Comparison between Japanese Wordnet and CSM data

4 Creation of New Knowledge System by Fusing Two Network Structure

In this section, we construct a conceptual system by fusing of semantic networks, i.e. Japanese Wordnet and set of word links extracted by CSM based method.

As CSM can extract many relations between words from the input data, i.e. huge corpus, we decided to set a threshold of CSM value to eliminate the number of links to fuse. We add links after elimination to Japanese Wordnet. There are 178,178 links stored in Japanese Wordnet as hyponym or hypernym. Among them, relations between nouns are 151,700. As we could get 151,604 relations with the threshold of 8200, we set the threshold 8200, and add these 151,604 relations to Japanese Wordnet, which means that we enlarge twice in size of conceptual system.

5 Human Association

5.1 Experiments by Human Subject

In order to verify our new concept system, we conducted experiments about human association with human subjects.

If a concept structure resembles to human knowledge structure, connecting two concepts in the concept structure means simulating human associations from one concept to the other. Wordnet resembles taxonomical knowledge that human made, and CSM data (NS data) shows the knowledge of scenes which humans picture in mind. Combining these two different kinds of knowledge, we are trying to create human knowledge structure which resembles more than other existing knowledge systems.

Figure 2. Experiments by high school pupils

We presented test sheet with 11 stimulus nouns to 51 participants (31 university students and 20 senior high school pupils), and asked them to write down what s/he associated by each stimulus words (Figure 2).

If we can find shorter connections of links between a stimulus word and an associated word via our system than Japanese Wordnet, our system is closer to knowledge structure of humans than Wordnet, at least from the viewpoint of associations by humans.

We made network of conceptual links between a stimulus word and associated words for visibility purpose.

For stimulus words, we use 11 nouns (music, curry, apple, soccer, scissors, communication, love, arm, pasta, school, vegetable), which were mostly selected by the following conditions;

Word stored in the Japanese Wordnet.
Concrete object.
Possibility of associations.
General word, not too specific.

With 11 stimulus words and 20 high school participants, we got a total of 1,456 words including 690 different words, such as "music: jazz, disco and rock" and "curry: rice, carrot."

5.2 Consideration to Simulate Associations

Figure 3 shows the association network using Pajek (Pajek—Program for large network analysis) for stimulus word "腕 ude 'arm' " by new concept system. Here association network means set of links between stimulus word and associated words. In Figure 3, only a stimulus word and associated words which are directly connected to a stimulus word or associated words are visualized in order to consider the relations between associated words.

There are "腕 ude 'arm' " and "アーム aamu 'arm' " in the Left-hand side of the figure. As both "腕 ude 'arm' " and "アーム aamu 'arm' " are in the same synset of Japanese Wordnet, this association, i.e. from "腕 ude 'arm' " to "アーム aamu 'arm' ", is a kind of paraphrase. The associated words shown in the middle of the figure are words directly associated from a stimulus words, such as "筋肉 kinniku 'muscle' ". There is "料理人 ryorinin 'cook' " which is different direction of association to other associated words. This association is caused by the polysemous feature of Japanese word "腕 ude 'arm' ", i.e. physical arm and ability about a technique. The word in the right-hand side of Figure 3, such as "タンパク質 tanpakushitsu 'protein' " can be thought as associations not directly from arm but via an associated word in the middle, e.g.,

"腕 ude 'arm' "
--- "筋肉 kinniku 'muscle' "
--- "タンパク質 tanpakushitsu 'protein' "

"腕 ude 'arm' "
--- "体 karada 'body' "
--- "服 fuku 'clothes' "

In order to make detailed discussion, we have to explore more about human association and network structure we created, however, current simple network seems to reflect procedure of human associations to a certain degree.

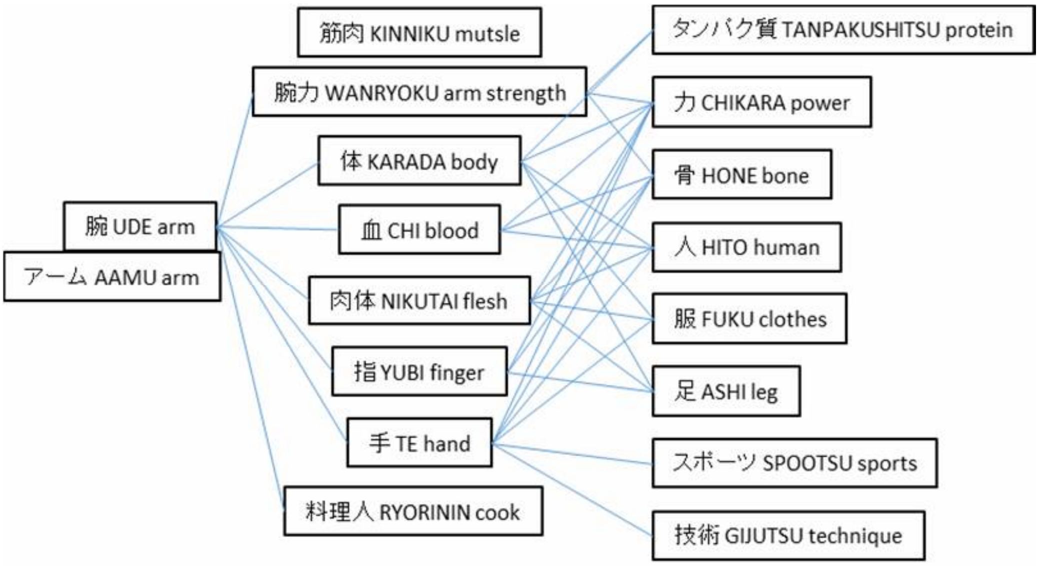

Figure 3. Partial association network for "腕 ude 'arm' "

6 Conclusion

In this study, we proposed and constructed a new conceptual system that by fusing of two different kinds of semantic networks. As semantic networks to fuse, we used the Japanese Wordnet which is a thesaurus made manually based on linguistic intuition and the new type of semantic structure which comes from the knowledge based on human utterances that are mutual relations of the concepts acquired automatically from the actual text.

In order to verify our new concept system, we conducted experiments by human subjects. We discussed the possibility of humanlike associations with our system.

We will consider tuning values assigned to each link of network precisely based on real associations conducted by humans by using simulation technology and huge computer power.

References

Norihiro Hagita and Minako Sawaki, 1995. Robust recognition of degraded machine-printed characters using complementary similarity measure and error-correction learning. *Proceedings of SPIE – The International Society for Optical Engineering,* 2442:236-244.

Hitoshi Isahara, Francis Bond, Kiyotaka Uchimoto, Masao Utiyama and Kyoko Kanzaki, 2008. Development of Japanese WordNet, LREC2008, Marrakech.

Daisuke Kawahara and Sadao Kurohashi, 2006. A Fully-lexicalized Probabilistic Model for Japanese Syntactic and Case Structure Analysis, *In Proceedings of HLT-NAACL2006*, 176-183.

Pajek—Program for large network analysis. Version 2.05. Available from: `http://vlado.fmf.uni-lj.si/pub/networks/pajek/`

Eiko Yamamoto, Kyoko Kanzaki and Hitoshi Isahara, 2005. Extraction of hierarchies based on inclusion of co-occurring words with frequency information. *Proceedings of IJCAI 2005*, 1166-1172.

Eiko Yamamoto and Hitoshi Isahara, 2011. Creative Information Retrieval Using Thematically Related Words. *Information Extraction from the Internet*, Chapter 13, iConcept Press