METHOD TO IDENTIFY DEEP CASES BASED ON RELATIONSHIPS BETWEEN NOUNS, VERBS, AND PARTICLES

Daisuke Ide[1] and Masaomi Kimura[2]
[1]Graduate School of Engineering and Science, Shibaura Institute of Technology
[2]Department of Information Science and Engineering, Shibaura Institute of Technology
3-5-7 Koto-ku Toyosu, Tokyo 135-8548, Japan

ABSTRACT

Deep cases representing the significant meaning of nouns in sentences play a crucial role in semantic analysis. However, a case tends to be manually identified because it requires understanding the meaning and relationships of words. To address this problem, we propose a method to predict deep cases by analyzing the relationship between nouns, verbs, and supplemental words, such as particles, in Japanese sentences. We also propose new deep cases based on a verb thesaurus and a deep case prediction method using a neural network.

KEYWORDS

Deep case; Case grammar; Text mining; Neural network; Clustering

1. INTRODUCTION

Cases represent the relationships between nouns and verbs in a sentence in Japanese. There are 2 types of cases, i.e., surface and deep cases. Surface cases classify noun roles obtained by the postpositional particle (particle), such as *ga*, *wo*, *ni*, *kara*, and *de*. Particles are similar to prepositions in English sentences. Typically, a particle is often located after a noun. Deep cases express the roles of words in a sentence (Ito, 2002, pp. 101-110). For example, in the sentence, *Tokyo-de asobu (play), de* is a particle denoting a surface case and its corresponding deep case is *location*. The deep case is important in sentence analysis because it reveals patterns of semantic relationships, which is difficult to accomplish with surface cases.

Semantic analysis of sentences is widely applied in many fields, such as machine translation and question answering. Therefore, automated semantic understanding is necessary for effective semantic analysis. However, cases tend to be manually identified because understanding the meaning and relationships of words is difficult for computer programs.

Shibuki et al. (2003) proposed a method to identify deep cases in sentences comprising a verb and 2 nouns (pp. 91-92). They reported a precision of 75.4%. Although some surface cases have corresponding deep cases, many surface cases, such as *ni, de,* and *wo*, can correspond to multiple deep cases (Shibuki et al., 2006, pp. 1413-1428). Takeno et al. (2014) proposed a method to identify deep cases that correspond to the *ni* case (pp. 1011-1014) and reported a precision of approximately 62%. However, it should be possible to apply deep case identification to any sentence because a sentence comprises an unspecified number of nouns and verbs. Moreover, to utilize such identification in natural language processing, generalization of deep case identification for all surface cases is required.

In our previous research (Ide and Kimura, 2016, p. 42), we proposed a method to identify the roles of particles considering the diverse correspondences between particles and their roles based on a previous study (Kimura, 2015, pp. 409-414). Moreover, we extracted relationships between each obtained role and deep cases proposed by Fillmore (1969).

In this study, based on a verb thesaurus (Inui et al., 2010), we define new deep cases wherein surface cases and deep cases are assigned to sentences and extract their characteristics. Furthermore, we propose a method using a neural network to identify the proposed deep cases of a target noun phrase (a noun followed by a particle) from the relationships between nouns, particles, and verbs based on the extracted characteristics. To apply this method to any sentence, we use a combination of a noun, particle, and verb, which is a minimum set carrying the meaning of the sentences.

2. PROPOSED METHOD

2.1 Classification of Deep Cases

A verb thesaurus has multiple fields such as verbs, nouns, particles, surface cases, and deep cases. In particular, many deep cases appear in the deep case field. In our previous study, we examined the relationships between particles and deep cases and found that deep cases can be partially identified by the particle (or the surface cases). Based on this idea, in this study, we classify deep cases in advance and identify new classified deep cases.

For example, the deep case *start point* is assigned to the noun *yane* (roof) in the sentence *yane* (roof)-*kara* (from) *ochiru* (fall down) and *before the change* is assigned to *Osaka*, a large city in Japan, in the sentence *Osaka*-*kara* (from) *itensuru* (transfer). Both sentences whose nouns refer to locations and verbs express movements have a similar meaning. Therefore, the deep cases of both sentences can be unified as a single deep case.

To extract the relationship between the surface cases and the deep cases in the verb thesaurus, we counted their co-occurrence frequency. Then, to classify the deep cases, we applied a hierarchical clustering algorithm to a deep case vector whose elements are the relative frequencies of surface cases appearing in target sentences in the verb thesaurus. Finally, we obtained new deep cases based on the characteristics of each cluster. We employed the *Ward method* to measure the distance between clusters.

2.2 Identification of Deep Cases

In some sentences, deep cases cannot be uniquely identified. For example, in the sentence *kuko* (airports)-*wo* (from and to) *hattyakusuru* (shuttle), *kuko* (airports) can be assigned to both start-point and end-point deep cases. To identify deep cases for such sentences, we employed a neural network to calculate the degrees of assignments to each deep case.

We used the flags of the deep cases as outputs of the neural network, as shown in Figure 1. After training, it was expected that we would obtain a continuous assignment degree value in the range [0, 1] for each deep case as the output of the neural network. Table 1 provides an example of the neural network output and lists each output value and its rank in descending order. In this example, deep case 2 obtained the highest degree and deep case 1 obtained the second highest degree. If multiple deep cases have high values, both can be assigned. In this study, we restricted the assignable number of deep cases to 2.

Numeric values were assigned to nouns, verbs, and particles. These values were used as input for the neural network. Similar to the output, noun, verb, and particle flags were used to obtain the numeric values. However, assigning a flag to each word is not desirable because compared to particles, there are many more similar noun/verb types. For example, consider the sentences *Tokyo he iku* (I will go to Tokyo) and *Osaka ni shuppatsusuru* (I will depart to Osaka). Obviously, *Tokyo* and *Osaka* are place names, i.e., nouns, and *iku* (go) and *shuppatusuru* (depart) are verbs expressing movement. Even though the words are not the same, the difference does not affect the assignment of deep cases to the nouns in the sentences. Thus, we must classify the nouns and verbs based on thesauruses in advance. The neural network input is illustrated in Figure 2.

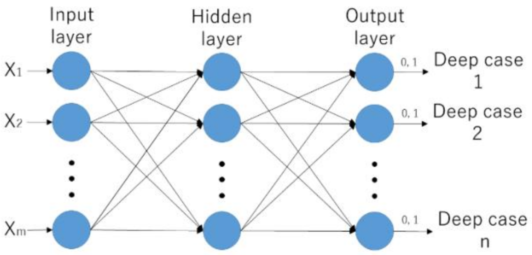

Figure 1. Neural Network Output

Table 1. Examples of Neural Network Output

Output	Output values	Descending order
Deep case 1	0.40	2
Deep case 2	0.70	1
…	…	…
Deep case n	0.03	N

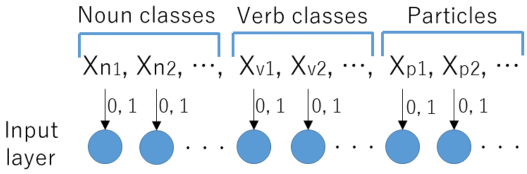

Figure 2. Neural Network Input

3. EVALUATION

3.1 Classification of Deep Cases

We extracted the frequency of surface cases corresponding to each deep case from the verb thesaurus and created a vector whose elements are their relative frequencies.

To organize the data, we unified the variants of expressions, such as reversed particles, e.g., *to-ni* and *ni-to*, in the surface case fields of the verb thesaurus.

The deep case categories were too granular; therefore, we unified obviously similar categories. For example, since *target (person)*, *target (biological)*, and *target (body part)* appeared in the deep case field, we unified them as *target*.

The dendrogram obtained via hierarchical clustering analysis applied to the frequency vectors is shown in Figure 3.

We set the cut-off threshold $\theta = 1$ to separate clusters and assigned numbers to clusters from left to right.

To extract the characteristics of each cluster, we extracted frequent nouns, verbs, and particles in deep cases classified to each cluster. We employed the database of classification vocabulary (DOCV) (National Institute for Japanese Language and Linguistics, 2004) and the verb thesaurus to classify the meaning of each noun and verb.

The frequent surface cases, verb classes, and noun classes of each cluster are shown in Tables 2, 3, and 4, respectively.

In Table 2, we find that the *ni*, *ga*, and *de* cases appear in multiple (but specific) clusters; however, the *kara*, *to*, *he*, *made*, and *ha* cases appear in only 1 cluster. Therefore, we can say that the surface and deep cases tend to demonstrate a relationship.

Table 3 shows that *position change* appears in multiple clusters and that *relationship* appears only in Cluster 7.

In Table 4, noun classes representing people appear in Clusters 5 and 7 and those representing location appear in Cluster 2.

Based on these observations, we summarize the characteristics of each cluster as follows:

- Cluster 1 represents the end point in action or state changes because it includes many particles, such as *ni*, *he*, and *made*, and verb classes that represent the changes.
- Cluster 2 represents the situation in action or state changes because it includes many noun classes representing locations and deep cases, such as time, location, and situation.
- Cluster 3 represents the affected thing in action or state changes because it includes many *wo* particles and the target of the deep case.
- Cluster 4 represents the means in action or state changes because it includes many noun classes, such as tool, material, and substance, and deep cases, such as tool and means.
- Cluster 5 represents the agent in action or state changes because it includes many noun classes representing the person and particles, such as *ga* and *ha*.
- Cluster 6 represents the original state in action or state changes because it includes many *kara* particles and deep cases, such as the start point and the original state.
- Cluster 7 represents the thing related to the target in action or state changes because it includes many verb classes representing relationship and deep cases, such as mutual and joint.

Based on these characteristics, we defined 7 deep case types: *Start point*, *End point*, *Situation*, *Target*, *Relationship*, *Agent*, and *Tool*.

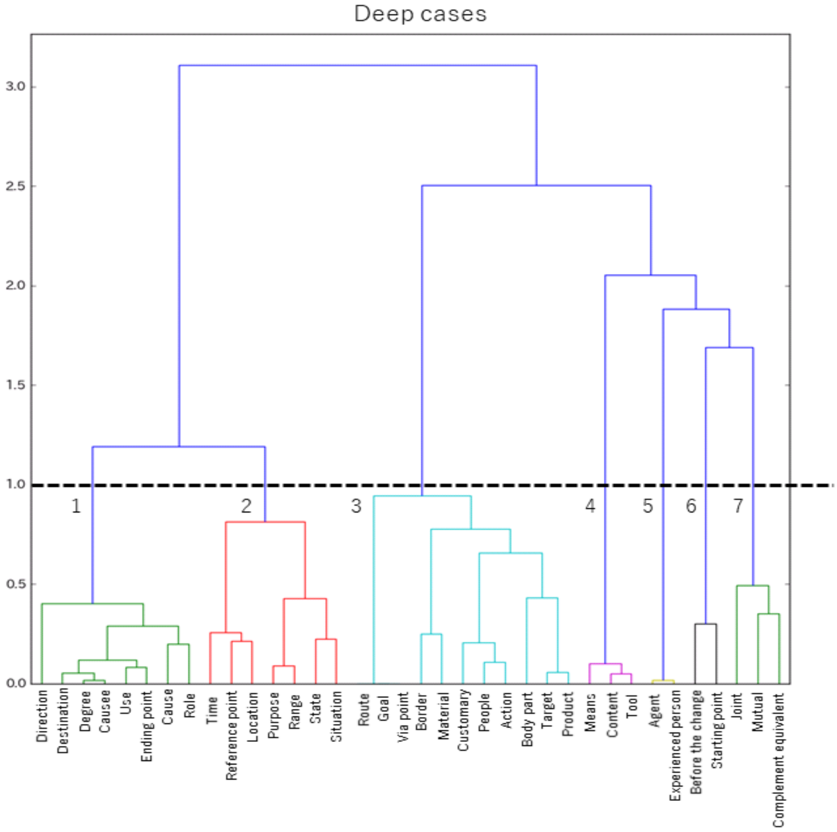

Figure 3. Deep Case Classification Results

Table 2. Frequent Surface Cases of Each Cluster

Cluster	Frequent surface case
1	*Ni, Ni–He, and Ni–Made*
2	*Ni, Wo, and De*
3	*Wo, Ga, and Ni*
4	*De and Ni*
5	*Ga, Ga–To, and Ha*
6	*Kara, Wo–Kara, Wo, and Ni–Kara*
7	*To, Ni, and Ni–To*

Table 3. Frequent verb classes of each cluster

Cluster	Frequent verb classes
1	Position change, Change of agent, Change of target, and Action of encourage the behavior to others
2	Creation–Annihilation, Moving action, Location, Physical behavior
3	Position change, Change of agent, Change of target, and Creation–Annihilation
4	Position change and Change in relationship
5	Position change, Change of agent, Creation–Annihilation, Action to person-object, Physical behavior, and Change of relationship
6	Position change and Change of agent
7	Change in relationship, Relationship with the other, Relationship, and Co-action

Table 4. Frequent noun classes of each cluster

Cluster	Frequent noun classes
1	Unregistered word, Members, Society, Housing, and Public and private
2	Unregistered word, Space, Society, Public and private, Residential, and Heaven and Earth
3	Unregistered word, Mind, Language, Volume, Body, and Life
4	Unregistered word, Tools, Substance, Materials, and Language
5	Members, Human, Family, Unregistered word, and Person
6	Unregistered word, Society, Housing, Space, Public and private, and Members
7	Members, Unregistered word, Companion, Family, and Human

3.2 Deep Case Identification

We evaluated the deep case identification of nouns using the proposed neural network.

We employed the DOCV and the verb thesaurus to classify the meanings of verbs and nouns. In addition, we prepared 5000 training data for each deep case. We oversampled the data for deep cases with fewer sample nouns, except for the nouns not registered in the DOCV. The number of original data for each deep case was 252 *Start point* data, 1427 *End point* data, 213 *Situation* data, 5442 *Target* data, 315 *Relationship* data, 3134 *Agent* data, and 77 *Tool* data.

We set 43 noun classes, 45 verb classes, and 12 particles as the inputs to our neural network, 7 neurons as a hidden layer (sigmoid layer), and 7 deep cases as outputs (linear layer). We used 35000 training data and trained by back propagation.

For evaluation, we calculated the precision of deep case identification with the neural network. The test data comprised 10860 samples, including 288 nouns in *Start point*, 1404 in *End point*, 213 in *Situation*, 5476 in *Target*, 289 in *Relationship*, 3096 in *Agent*, and 94 in *Tool*. We also calculated the precision of a deep case with the second highest output value for samples that were incorrectly identified, which comprised 1408 data, including 22 in *Start point*, 118 in *End point*, 59 in *Situation*, 989 in *Target*, 30 in *Relationship*, 176 in *Agent*, and 14 in *Tool*. The results are shown in Figure 4. The left bars in the graph show the precision of the deep case with the highest output value, whereas the right bars show the precision with the second highest output value.

In each deep case, the precisions were 92.4% for *Start point*, 91.6% for *End point*, 72.3% for *Situation*, 82.0% for *Target*, 89.6% for *Relationship*, 94.3% for *Agent*, and 85.1% for *Tool*. The precisions with the second highest output were greater than 90% compared to the test data that was incorrectly identified for the deep case other than *Agent* and *Target*.

It is evident that *Agent* and *Target* achieved high precision. This could be because the neural network was well trained for them since the number of original training data was sufficiently large.

Situation (shortest left bar in Figure 4) also has the shortest right bar whose precision was 21.6%. The features could not be obtained easily because all frequent surface cases in *Situation* also appeared in other deep cases, as shown summarized Table 2. In addition, the neural network could not be trained sufficiently because there were too few training data.

However, the overall precision of the neural network was up to 87%, which is higher than the respective precisions reported in previous studies (75.4% and 62.0%) (Shibuki et al., 2003; Takeno et al.,2014).

Then, to observe the effect of the role of noun classes, we used data that included nouns that are not registered in the DOCV. We evaluated the neural network with 2115 test data, including 70 in *Start point*, 331 in *End point*, 65 in *Situation*, 935 in *Target*, 102 in *Relationship*, 589 in *Agent*, and 23 in *Tool*. To input the noun values, we set all flags to 0. Again, we calculated precision for deep cases with the second highest output value for 516 test data, including 17 in *Start point*, 22 in *End point*, 31 in *Situation*, 185 in *Target*, 12 in *Relationship*, 242 in *Agent*, and 7 in *Tool*.

The results are shown in Figure 5. The precisions of *Target*, *End point*, and *Relationship* are close to those shown in Figure 4. Therefore, their deep cases are less affected by nouns because they can be identified only by a set of particles and verbs. The *Relationship* deep cases can be identified independent of the nouns because there are specific particles and verbs paired to *Relationship*, as summarized in in Tables 2 and 3.

In contrast, *Agent*, *Start point*, and *Situation* were significantly affected by nouns because their precision greatly decreased compared to the results shown in Figure 4. However, their precisions for the deep case of the second highest output value increased. This suggests that although it is difficult to correctly identify the deep cases without nouns, we can the limit candidates of a correct deep case.

These results suggest that we can apply the proposed neural network to identify the deep cases of pronouns.

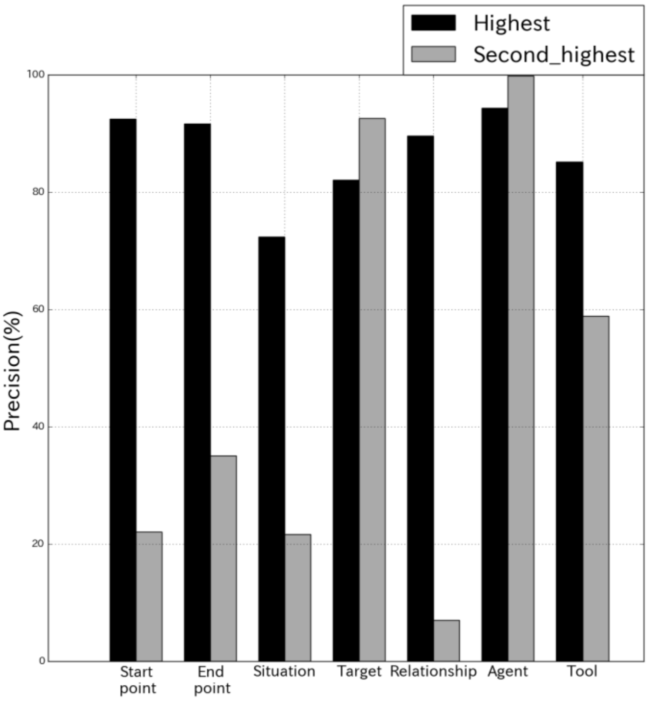

Figure 4. Precisions of neural network obtained with all training data

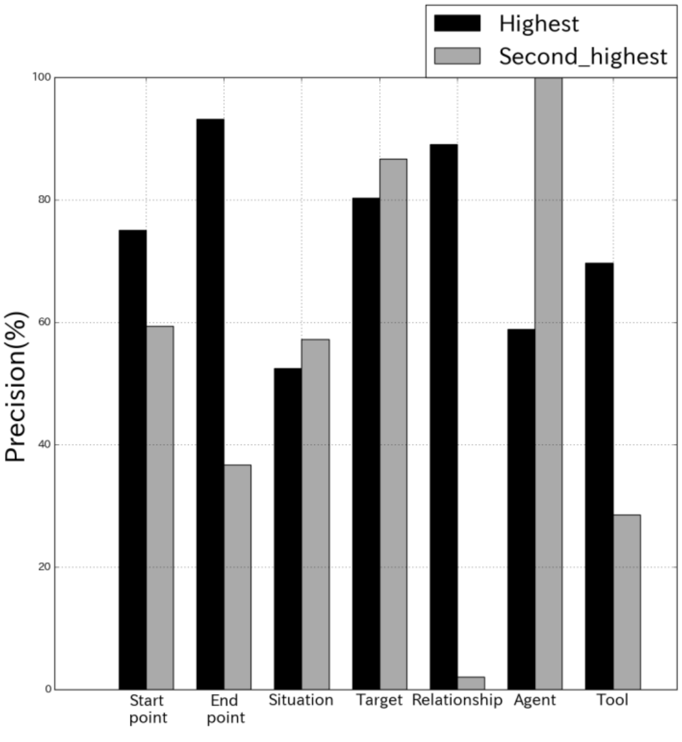

Figure 5. Precisions of neural network obtained using training data with unregistered nouns

4. CONCLUSION

In this study, we proposed a method to identify deep cases from the relationship between nouns, verbs, and particles in order to mitigate the problem by which deep cases are required to be manually identified.

We classified deep cases based on the co-occurrence frequency between the deep cases and surface cases in a verb thesaurus by applying a hierarchical clustering algorithm, and we have defined 7 new deep case types: *Start point*, *End point*, *Situation*, *Target*, *Relationship*, *Agent*, and *Tool*.

Then, to identify the proposed deep cases for sentences, we employed a neural network to calculate the degrees of assignments to each deep case. We used the flags of noun classes, verb classes, and particles as input to the neural network and flags of the proposed deep cases as output of the neural network, and we evaluated the trained neural network using test data.

For each deep case, the precisions were 92.4% for *Start point*, 91.6% for *End point*, 72.3% for *Situation*, 82.0% for *Target*, 89.6% for *Relationship*, 94.3% for *Agent*, and 85.1% for *Tool*. The overall precision of the neural network was up to 87%. The precisions with the second highest output were greater than 90% compared to test data that were incorrectly identified for the deep case other than *Agent* and *Target*. *Target*, *End point*, and *Relationship* were less affected by nouns. In contrast, *Agent*, *Start point*, and *Situation* were significantly affected by nouns.

In the future, we will increase the volume of training data to improve the precision of the neural network and plan to extend a method to estimate omitted words, such as pronouns or zero pronouns, based on the deep cases identified in this study.

ACKNOWLEDGEMENT

In this study, we used the verb thesaurus created by Koichi Takeuchi, Kentarou Inui, Nao Takeuchi, and Atsushi Fujita with the support of the Ministry of Education, Culture, Sports, Science and Technology Grant-in-Aid for Scientific Research (B), “Application for engineering realization and paraphrase knowledge acquisition of paraphrase calculation mechanism based on lexical semantics” (17300047, representative: Kentaro Inui).

REFERENCES

Ito, K. (2002). About semantic interpretation of case particle representation in Japanese. *Meikai Japanese, 7*, 101-110.

Shibuki, E., Araki, K., & Tochinai, K. (2003). Automatic guess technique of deep cases based on the principles and rules of one case per a sentence. *Proceedings of Forum on Information Technology, 2*, 91-92.

Shibuki, E., Araki, K., Momouchi, Y., & Tochinai, K. (2006). A method for inference of deep case based on deep case preference of word concept. *The IEICE Transactions on Information and Systems (Japanese Edition), 6*, 1413-1428.

Takeno, S., Matsuda, M., Kajiwara, T., & Yamamoto, K. (2014). Study of the automatic grant of ni case of deep case using machine learning. *Proceedings of The Association for Natural Language Processing, 20*, 1011-1014.

Ide, D., & Kimura, M. (2016). Method to estimate deep cases based on relationship between particles and verbs. *Proceedings of the IEICE General Conference, D-5-3*, 42.

Kimura, M. (2015). A proposal of topic map based dialog system (2nd report). *Proceedings of Fuzzy System Symposium, 31*, 409-414.

Fillmore, C. J. (1969). *The case for case. In Bach and Harms (Ed.) universals in linguistic theory*. New York: Holt, Rinehart, and Winston.

Inui, K., Takeuchi, K., Takeuchi, N., and Fujita, A. (2010). Application for engineering realization and paraphrase knowledge acquisition of paraphrase calculation mechanism based on lexical semantics. *Technical report, Ministry of Education, Culture, Sports, Science and Technology Grant-in-Aid for Scientific Research (B).*

National Institute for Japanese Language and Linguistics (2004). *Enlarged and revised edition database of classification vocabulary*. Japan: Dainippon Tosho