ESTIMATION OF SEMANTIC CASE OF JAPANESE DIALOGUE
BY USE OF DISTANCE DERIVED FROM STATISTICS OF DEPENDENCY

Tomonobu Saito¹ and Kiyoshi Hashimoto²

¹ AI-KC Laboratory, Graduate School of IS, The University of Electro-Communications, Chofu-shi, Tokyo, 182-8585 JAPAN
² Hashimoto Research Laboratory 5-13-10 Takanodai, Nerima-ku, Tokyo, 177-0033 JAPAN

ABSTRACT

In an attempt to estimate the semantic cases for noun-particle-verb triples in the ATR dialogue corpus, the authors propose a measure of distance based on statistics of dependent noun-particle-verb triples. A clustering analysis of all the triples in the corpus was conducted using the measure of distance. Competence of the proposed measure of distance is verified by examination of the distribution of the single-case clusters. By use of the score derived from the measure of distance of the training corpus, the authors conducted the estimation of the correct semantic case for a given noun-particle-verb triple in the test corpus. The result remarkably differentiates the particles with respect to the estimation accuracies. For instance, particle ‘wo’ has accuracies over 80 %, while ‘de’ has accuracies less than 40 %. The correlation analysis between the accuracy and the consistency have also tendencies to higher accuracies.

1. INTRODUCTION

As a preliminary step to acquire the correct case-frame tree from Japanese dialogue sentence, we previously made a connectionist approach in estimation of the semantic case for a given 135 sentences in the ATR dialogue corpus [1,2]. Accuracy for 23 cases achieved by means of neural network alone was 62 %, while joint use of IPAL dictionary and neural network improved the accuracy up to 65 %. In this study, we make a new statistical approach in estimation of 64 semantic cases for given noun-particle-verb triples in the test corpus by use of distance derived from statistics of dependency.

2. SEMANTIC CASE

The ATR corpus contains sentences of dialogue related with an information office for an international conference [3]. In most previous studies related with ‘case-frame’, numbers of the total set of cases were less than 20. However, in this study, all of the 64 cases that are listed as the semantic tags in the corpus is used in order to obtain precise semantic information in the sentences. The 64 semantic tags includes precise tags such as ACC (accompany), AGT (agent), AVO (value-of-object), CAU (cause, reason) CIR (circumstance) CNC (concession), EVA (evaluation), ORG (organizer), SPA (space, all), SPT (space, terminal), SRC (surroundings change), TMA (time, all), TMT (time, terminal), UVS (unvoiced subject) and REL (relation).

3. STATISTICAL DISTANCE

As the first step to the statistical measure for our study, we define the distance based on the frequency of dependent noun-particle-verb triples. Later, the estimation of the correct semantic case for a given noun-particle-verb triple is conducted simply by searching for the case that has the maximum score. The score will later be hierarchically derived from this statistical distance.

3.1 Distance between Nouns

Similarity between two nouns $n_1$, $n_2$ can be well represented by the correlation coefficient around origin, or called ‘cosine’, between $f_{1i}(v_i)$ and $f_{2i}(v_i)$

$$r(n_1, n_2) = \frac{\sum f_{1i}(v_i) f_{2i}(v_i)}{\sqrt{\sum f_{1i}(v_i)^2 \sum f_{2i}(v_i)^2}}, \quad (1)$$

assumes 1 when $f_{1i}(v_i)$ is identical to $f_{2i}(v_i)$, and $\theta$ when $f_{1i}(v_i)$ is not correlated with $f_{2i}(v_i)$. The similarity $(n_1, n_2)$ is used to define the distance between $n_1$ and $n_2$.

3.2 Distance between Verbs

$$d(n_1, n_2) = \frac{1}{r(n_1, n_2)} - 1. \quad (2)$$

Similarity and distance between two verbs $v_1$, $v_2$ can also be defined essentially in the same way as equations (1) and (2), based on the statistics of frequencies $f_{1i}(n_i)$ and $f_{2i}(n_i)$ of dependency pairs $(n_i, v_i)$ and $(n_i, v_j)$. In other words, the similarity $(v_1, v_2)$ is obtained as the correlation coefficient by equation (1) and the distance $d(v_1, v_2)$ is calculated by the similar equation as (2).

3.3 Distance between Two Noun Verb Pairs

As the Euclidean sum of both distances between nouns $n_1$ and $n_2$ and between verbs $v_1$ and $v_2$, we obtain the distance between a noun-verb pair $n_1v_1$ and another noun-verb pair $n_1v_2$.

Since the distance thus derived is statistics of a whole corpus, inadequate statistics due to the sparse noun-verb pair can be avoided. Table shows four examples of distances between two noun-verb pairs. Each distance depends on the noun-verb pair regardless of the case particle between the noun-verb pair.

Table Examples for distance between two noun-verb pairs.
4 CONSISTENCY OF CLUSTER WITH CASE

4.1 Clustering of Noun-Verb Pairs

The first experiment is clustering analysis of all the triples in the corpus using the proposed measure of distance. The purpose of clustering analysis is verification of the competence of the proposed measure of distance by way of examination of the distribution of the cases over the corpus. We adopted ‘greedy algorithm’ as the clustering algorithm in this study. It is composed of the following four steps:

1. to assign an initial cluster to all the N noun-verb pairs,
2. to merge two minimum distance clusters into a single cluster,
3. to repeat step (2) C times, and
4. to stop clustering if all the inter-cluster distances are infinite or all the clusters reduces to a single cluster.

In step (4), we use the average of the distance of all the combinations of the two clusters as the inter-cluster or mutual distance. All of the 766 noun-particle-verb triples in the ATR dialogue corpus were divided into 12 groups with respect to the case particles. Application of the clustering algorithm to each group of noun-particle-verb triples produces Nc clusters.

4.2 Consistency of Cluster with Case

A cluster that contains more than one semantic case is semantically ambiguous because of its multiplicity of meaning. If clustering process produces more single-case clusters in each group of the corpus, the group is thought to have that much more consistency of cluster with case. Since the corpus were divided into 12 groups with respect to the case particles, step by step tracing of the clustering processes reveals the variety in existence rate of single-case clusters in each particle group. Application of the clustering algorithm to each group of noun-particle-verb triples produces Nc clusters.

4.3 Experimental Result of Clustering

The rate for case consistency is updated every time the clusters are merged. In this study, we attempt to make two kinds, namely A and B, of graphical representations for case consistency. The first kind, case consistency rates A, is those in terms of the distance at the time of merge, namely, cluster size.

4.3.1 Case Consistency Rate A

Fig. 1 shows case consistency rates for six high frequency case particles as a function of cluster size. Fig. 2 shows case consistency rates for six low frequency case particles as functions of cluster size. A general survey of these two figures indicates a remarkable variety of consistency rates among 12 case particle groups. Since the merging process terminates at sizes more than 2.5, the consistency rates, for instance, at size 2.0, elucidates the variety of consistency rates with respect to case particles. The cross section at size 2.0 of both figures indicates that all the case consistency rates for six low frequency case particles exceeds 90% and that the consistency rates for six high frequency case particles dispersed with the minimum rate 46 % for particle ‘to’. Scrutiny of Fig. 1 reveals that the case consistency rates for particles ‘ga’ and ‘wo’ are high in wide range of cluster size and that, on the contrary, consistency rate for particle ‘to’ rapidly decreases as the cluster size increases. The consistency rates for particles ‘kara’, ‘de’ and ‘ni’ are intermediate in between two extremities.

4.3.2 Case Consistency Rate B
The second attempt of graphical representations for the case consistency B is made in terms of normalized number of clusters. Figs. 3 and 4 show the case consistency rates for six high frequency case particles and for six low frequency case particles, respectively, each as functions of normalized number of clusters.

![Graph 1: Case consistency rate vs. normalized number of clusters for six high frequency case particles.](image1)

![Graph 2: Case consistency rate vs. normalized number of clusters for six low frequency case particles.](image2)

A survey of these two figures reveals that the termination of the merge occurs at a variety of normalized number of clusters less than 60% and that the case consistency rates reduce rapidly at a variety of normalized number of clusters less than 40% ~ 30% due to the increase of ambiguity of the particles by forced merge of clusters. A description will be later given in this study on a statistical analysis of the relation between the case consistency rates and the accuracies of estimation of semantic case.

5 ESTIMATION OF SEMANTIC CASE

5.1 Case Score

The principal aim of this study is the estimation of the correct semantic case for a given noun-particle-verb triple. The score necessary for the estimation is calculated with respect to the triple. For a given triple \((n, k, v)\) in the training samples, the case score for estimation of a semantic case is calculated by the criterion function defined by

\[
S(n, k, v, \alpha) = \sum r(n, k, v, \alpha)
\]

where the partial case score \(r(n, k, v, \alpha)\) in Eq.(6) is given by

\[
r(n, k, v, \alpha) = \frac{1}{d(<n,v>,<n_i,v_i>)+1}
\]

the distance between a data pair in the data pair set,

\[
D_T(n, k, v, \alpha), D_T(n_i, k, v_i, \alpha)
\]

Note that \(r(n, k, v, \alpha)\) is equal to zero if the training data pair set is empty. Thus, in the training process, each of the noun-particle-verb triples is given the score \(S(n, k, v, \alpha)\). At the estimation stage, the score necessary for the decision of the case is the score \(S(n, k, v, \alpha)\) given at the training process.

5.2 Experiment of Estimation of Semantic Case

We conducted the experiments of estimation of the semantic case of the noun-particle-verb triples by splitting a whole data into the training and the test data. The distribution of the score with respect to obtained in the training processes are applied to the estimation of the correct semantic case of noun-particle-verb triples in the test set sentences. The ratio of the training data to the test data is varied from 1:9 to 9:1. For an experimental condition with each of the ratio, the training and the test processes are repeated five times with the randomly selected data. The mean value is used as the result.

![Graph 3: Accuracies for case estimation for high frequency case particle.](image3)

![Graph 4: Accuracies for case estimation for low frequency case particle.](image4)
and for low frequency case particle, respectively. Fig. 5 indicates that, for high-frequency case particles, variations of the case estimation accuracies are less than 30% with respect to the ratio of test data and that the variations are remarkable with respect to particles. For instance, particle 'wo' shows accuracies over 80%, while 'de' shows accuracies less than 40%. Fig. 6 indicates that the variations of the accuracy for low-frequency case particles are much more remarkable than those for high frequency case particles and that those accuracies tend to decrease as the ratio of test data increases.

5.3 Correlation Analysis of Experimental Results

In order to search for factors having an effect on the accuracies for case estimation, we conducted a correlation analysis among the sample size of each case particle, case consistency rates and estimation accuracies. For the correlation analysis, we adopted the case consistency rates at cluster size 2 and at normalized number of clusters 60% as the representative consistency rates A and B, respectively. We also adopted the highest values for case estimation accuracies as data for correlation analysis.

Table 3 shows the sample size of each case particle, case consistency rates A and B and estimation accuracies. Table 4 shows the results of correlation analysis. The correlation coefficient 0.57 between the accuracy and the consistency rates B suggests that the particles of higher accuracies are apt to have higher consistencies, while the correlation coefficient -0.32 between the accuracy and sample size indicates their feebly inversive or compensatory relation.

Table 4 Correlation among frequencies of case particles, case consistency rates A and B and estimation rates of semantic cases.

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>Consistency A</th>
<th>Consistency B</th>
<th>Accuracy R</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.00</td>
<td>1.00</td>
<td>-0.09</td>
<td>1.00</td>
</tr>
<tr>
<td>Consistency B</td>
<td>0.41</td>
<td>0.57</td>
<td>1.00</td>
</tr>
</tbody>
</table>

CONCLUSIONS

We made a new statistical approach in estimation of semantic cases in a dialogue corpus by use of score derived from statistics of dependency. Since the distance thus derived is statistics of a whole corpus, inadequate statistics not only due to the sparse noun-verb pair but also due to the sparse noun-particle-verb triples can be avoided. However, the experimental result of estimation of semantic case is not always satisfactory. Analyses of the estimation errors will suggest the effective method for improvement of the estimation accuracy. Also the similar experiments by reduced number of the cases from current 64 cases will indicate the optimal number of the cases. However, the next and the most important step of our study is making every effort to secure the correct case-frame tree as the building blocks of the results of current single case estimation.

REFERENCES