Statistical Matching of Two Ontologies

Satoshi Sekine **Kiyoshi Sudo**
Computer Science Department
New York University
715 Broadway, 7th floor
New York, NY 10003 USA
[sekine|kiyo7793]@cs nyu edu

Takano Ogino
Electronic Dictionary Research
78-1 Sakumahigan, Kanda, Chioda-ku
Tokyo, 101-0026 Japan
ogino@edr co jp

1 Introduction

Standardizing ontologies is a challenging task Ontologies have been created based on different backgrounds, different purposes and different people However, standardizing them is useful not only for applications, such as Machine Translation and Information Retrieval, but also to improve the ontologies themselves During the process of standardization, people can find bugs or gaps in ontologies So standardization brings benefits compared to just using them separately There is a committee for standardizing ontologies at ANSI, the "ANSI Ad-Hoc Group for Ontology Standards" (Hovy 1996)

Although there have been a few attempts to merge and compare ontologies, this work is still at a preliminary stage of research (Ogino et al 1997) attempts manual merging of EDR (EDR 1996) (Miyoshi et al 1996) and WordNet (Wordnet) (Miller 1995), (Utiyama and Hashida 1997) used statistical methods to merge EDR and WordNet (Pangloss) is also working on standardizing ontologies It is certain that manual methods have great difficulty in matching the entire ontologies It would require three thousand years for a person to check all possible node pairings, if the two ontologies have 40 000 nodes each and each judgement takes a minute So automatic methods are needed to find matches automatically or at least to narrow down the candidates for matching

In this paper, we investigate a simple statistical method for matching two ontologies The method can apply to any ontologies which are formulated from is-a relationships In our experiments, we used EDR and WordNet This work is similar to the work in (Utiyama and Hashida 1997) They defined the task as the MWM (Maximum Weight Match) of bipartite graphs, an approach which is basically common to most ontology matching schemes The information they used is partially fuzzy, i e for calculating the distance between two nodes, they used the information from each node and its neighborhood, not distinguishing between information from parent and child nodes However, since the structure of the ontologies (the relation between parent and children) is significant, it might be better to utilize such structural information In our experiments, we will focus on this issue, rather than trying to achieve a higher performance The importance of parent, child and grandchild information will be examined We will conduct several experiments with or without some of the information It is also important to discover what weighting balance gives good matches

2 Ontologies

First we will briefly explain the ontologies we used in our experiments

2.1 EDR

The EDR Concept Dictionary contains 400,000 concepts listed in the Japanese and English Word Dictionaries of 200,000 words each The EDR Concept Dictionary is one of the five types of EDR dictionaries, the others are the Word Dictionaries for English and Japanese the Bilingual Dictionary, the Coocurrence Dictionary, and the Technical Terminology Dictionary The EDR Concept Dictionary consists of three sub-dictionaries the Headconcept Dictionary contains concept explanations in natural language (both in English and Japanese), the Concept Classification Dictionary contains a set of is-a relationships, and the Concept Description Dictionary contains pairs of concepts that have certain semantic relationships other than is-a relationship i e *object*, *agent goal*, *implement a-object* (object of a particular attribute), *place*, *scene* and *cause*

The Concept Classification Dictionary classifies all the 400 000 concepts based on their meaning A polysemous word is put into several word classifications (concepts) As multiple inheritance is allowed, the entire structure is not a tree but a DAG (directed acyclic graph) There are 6,000 intermediate nodes and the maximum depth is 16

2 2 WordNet

WordNet (Wordnet) is an English ontology The nodes are represented by a set of synonym words (called 'synsets') WordNet contains 60,557 noun

synsets, 11,363 adjective synsets, and 3,243 adverb synsets Between synsets, there are relations which include (but are not limited to) hypernymy/hyponymy, antonymy, entailment and meronymy/holonymy A word or collocation may appear in more than one synset, and in more than one part of speech The words in a synset are logically grouped such that they are interchangeable in some context

3 Experiments

The basic idea of the matching is to find the distance (similarity) between a node in EDR and a node in WordNet There could be several strategies for defining a distance between two nodes, we will use the words attached to each node and its parent, child and grandchild in the computation We did not use the descriptions of concepts

As a preliminary experiment, we restricted the number of nodes to be considered, because both ontologies are big We used the nodes at the top 5 levels (distance from the top is at most 5) and deleted nodes which have no English words and no descendents in EDR (some EDR nodes have only Japanese words) This left 14,712 nodes in EDR and 5,185 in WordNet Even with these restriction, the number of possible pairings is 76,281,720 Our target is to find good matches among them

3 1 Definition of Distance

The distance between nodes is defined based on the notion which is commonly used, the dice coefficient Assume the node N_1 in ontology1 has n_1 words and N_2 in ontology2 has n_2 words, and there are m words in common The dice coefficient (DC) is defined as follows

$$DC(N_1, N_2) = \frac{2m}{n_1 + n_2}$$

Now we define the basic distance as 1 minus the value The smaller the distance, the closer the two nodes

$$dist(N_1, N_2) = 1 - \frac{2m}{n_1 + n_2} \quad (1)$$

We now define the distance of two nodes (N_1,N_2) based on the basic distance definition The words in parents, children and grandchildren are also used Such nodes are taken as a bag of nodes, i e only one set of words is created for each category regardless of the number of nodes Such a bag of nodes is represented as N^{parent} and so on The distance of each category is calculated just like the basic distance In the following equation, *cat* should be replaced by *parent*, *itself*, *child* and *gchild* (for grandchild)

$$dist^{cat}(N_1, N_2) = dist(N_1^{cat}, N_2^{cat}) = 1 - \frac{2m^{cat}}{n_1^{cat} + n_2^{cat}}$$

Then interpolation is used to merge the four basic distances in order to keep the range from 0 to 1 We introduce four coefficients c^{parent},c^{itself},c^{child},c^{gchild} to define the node distance , $D(N_1, N_2)$

$$\begin{aligned} D(N_1, N_2) = {} & c^{parent}\ dist^{parent}(N_1, N_2) + \\ & c^{itself}\ dist^{itself}(N_1, N_2) + \\ & c^{child}\ dist^{child}(N_1, N_2) + \\ & c^{gchild}\ dist^{gchild}(N_1, N_2) \end{aligned}$$

$$c^{parent} + c^{itself} + c^{child} + c^{gchild} = 1 \quad (2)$$

The coefficients (c^{cat}'s) will be the important factor in the experiments As will be described in the next section, we use several combinations of the coefficients to observe which information is important

3 2 Experiments

We conducted eight experiments using different combinations of the coefficients The first experiment uses only the information in the nodes themselves, while other experiments use the node and parent, the node and children, or all four sets Table 1 shows the coefficient combinations used in the experiments

Exp	parents	self	child	gchild
1	0 0	1 0	0 0	0 0
2	0 0	0 3	0 7	0 0
3	0 0	0 5	0 5	0 0
4	0 0	0 7	0 3	0 0
5	0 3	0 7	0 0	0 0
6	0 2	0 5	0 3	0 0
7	0 2	0 6	0 2	0 0
8	0 2	0 5	0 2	0 1

Table 1 Coefficient Combination

3 2 1 Analysis of the statistical results

Before describing the evaluation results, some interesting analyses are presented in this section These analyses do not concern directly the evaluation of the experiment, but indicate the nature of the experiments or the nature of the ontologies

Number of outputs

We used a threshold to restrict the number of outputs If the distance is greater than 0 9, the result is not generated Table 2 shows the number of outputs in each experiment Recall that there are 76,281,720 possible pairings of nodes It is interesting to see that the numbers are almost the same The number of outputs in Experiment-4 is slightly smaller, we believe this is because the weight assigned to the nodes themselves, which gives the greatest contribution, is low

Experiment	(Coefficients)	Output
1	(0 0, 1 0, 0 0, 0 0)	10,275
2	(0 0, 0 7, 0 3, 0 0)	10,151
3	(0 0, 0 5, 0 5, 0 0)	10,151
4	(0 0, 0 3, 0 7, 0 0)	9,093
5	(0 3, 0 7, 0 0, 0 0)	10,799
6	(0 2, 0 5, 0 3, 0 0)	10,098
7	(0 2, 0 6, 0 2, 0 0)	10,206
8	(0 2, 0 5, 0 2, 0 1)	10,028

Table 2 Number of Outputs

The numbers are around 10,000, which represents 0 013% of the possible matches This suggests that there is a possibility of narrowing down the matches to be examined by a human, as the distance 0 9 is very large and the number of outputs is so small To prove this assumption, we have to conduct an evaluation to see if there are good matches which were not generated This is beyond the evaluation in this paper, because it requires manual matching from scratch We will discuss this later

Complete Match

We can find the number of complete matches (which have exactly the same word(s)) by counting the pairs with distance 0 0 in Experiment-1 The number of complete matches is 1778, which is quite large compared to the number of nodes under consideration in WordNet (about 5,000) Also, by counting up the number of pairs with distance 0 0 in Experiment-5, we can find parent-complete matches which are complete matches where the parents also have the same words The number of parent-complete matches is 1 This is surprisingly small, even considering that we used only subsets of the ontologies The only parent-match is the following

```
parent   invertebrate
child    arthropod
```

Naturally people might guess that there would be more parent-complete matches For example, the name of a mammal might be a plausible candidate (where the parent is "mammal" and child is, for example, "elephant") However, this is not the case "Elephant" and "mammal" appear as follows (unrelated nodes are not shown)

```
EDR
<no English word, Japanese=mammal>
    +-----  <mammal, J-Description -
    |         an instance of mammal>
    +-----  <elephant>
```

WordNet

```
<mammal>
    +-----<proboscidean,proboscidian>
                  + ------ <elephant>
```

This is one of the typical problems of ontology design, how detail concepts should be introcuced Also, there is a translation problem in EDR, i e sometimes there is words or a description in only one language

There are some other reasons why the number of parent-matches is so small

- Some nodes in EDR have no words associated with them This is how the EDR Classification Dictionary was designed It is based on the classification of words into some predefined boxes, and not creating hierarchy of words It would be better to use the concept descriptions of the dictionary, although it is not clear how to compare a synset (set of words) and a description Also, we might be able to use information written in Japanese when there is no English word but there are Japanese words
- WordNet uses a synset to represent a node, whereas EDR's node is primarily represented by a description, there could be differences caused by this The average numbers of words in a node are also different

There were no children-matches, which are complete matches where the words in the child nodes are also the same The closest matches in Experiment-2 and 3 are the following

```
EDR
  parent(*)  year
  children   school year

WordNet
  parent(*)  year
  children  anomalistic year, lunar
     year, school year, academic year,
     solar year, tropical year, astro-
     nomical year, equinoctial year
   (There are actually 4 child nodes )
```

3 2 2 Evaluation

As it is impossible to evaluate all the results, we selected four ranges (rank 1 to 20, 501 to 520, 2001 to 2020, and 9001 to 9020) and the data in these ranges was evaluated manually Evaluation was done by putting the matches into three categories

- A Two nodes are completely the same concept
- B Other than A and C
- C Two nodes are completely different concepts

Category B includes several different things, including partial matches and ambiguous cases by the manual evaluation However, the number of results in this category was not so large, so it should not affect the overall evaluation Table 3 shows the evaluation result The columns represent the four ranges and the each row represents one of the eight experiments An element has

Experiment	1-20	501-520	2001-2020	9001-9020
1 (0 0, 1 0, 0 0, 0 0)	3 / 1 / 16	8 / 1 / 11	4 / 2 / 14	5 / 4 / 11
2 (0 0, 0 7, 0 3, 0 0)	6 / 1 / 13	6 / 1 / 13	3 / 3 / 14	1 / 2 / 17
3 (0 0, 0 5, 0 5, 0 0)	6 / 1 / 13	6 / 1 / 13	3 / 3 / 14	1 / 2 / 17
4 (0 0, 0 3, 0 7, 0 0)	2 / 1 / 17	10 / 3 / 7	4 / 4 / 12	5 / 5 / 10
5 (0 3, 0 7, 0 0, 0 0)	10 / 1 / 9	7 / 1 / 12	2 / 3 / 15	6 / 5 / 9
6 (0 2, 0 5, 0 3, 0 0)	11 / 1 / 8	6 / 1 / 13	2 / 3 / 15	5 / 9 / 6
7 (0 2, 0 6, 0 2, 0 0)	11 / 1 / 8	6 / 1 / 13	2 / 3 / 15	1 / 7 / 12
8 (0 2, 0 5, 0 2, 0 1)	11 / 1 / 8	6 / 1 / 13	2 / 3 / 15	5 / 6 / 9

Table 3 Evaluation Result

three numbers, corresponding to the categories A, B and C, separated by "/" We can't make a direct comparison to other methods For example, while (Utiyama and Hashida 1997) also used EDR and WordNet, they used only connected components and we imposed the level restriction However, relative comparisons among our 8 experiments are meaningful and important We will discuss them in the next section

3 3 Discussion

Using only the nodes themselves (Exp-1)

In Experiment-1, only the words in the nodes being compared are used The evaluation result was not very good For example, there are only 3 matches of category A in the highest range Based on an examination of the results, we observed that this is due to word polysemy Even if two nodes have a word in common, the word could have several meanings, and hence the corresponding nodes could have different meanings For example, the word "love" can mean "emotion" or "no point in tennis" To see how the results we obtained might arise, suppose a word has 4 senses in ontology1 and 5 in ontology2, and there are 3 senses which are the same in the two ontologies Then there are 20 pairings of the senses and out of them only 3 can be judged as category A Although this is just an assumption, the reality might not be that far from this explanation based on the observation of the result

Adding child nodes (Exp-2,3,4)

In Experiment-2,3 and 4, we used the information of the nodes themselves and their child nodes The evaluation results for Experiment-2 and 3 are the same, both of them have 6 A's in the highest range The number is twice that in Experiment-1 This improvement is due to disambiguation of polysemous words For example, the same sense of a polysemous word might have similar words in the child nodes, whereas it might be rare that different senses have the same words in the two ontologies

In Experiment-4, we put more weight on child nodes rather than the nodes themselves This experiment was conducted based on the assumption that the number of words in child nodes may be much larger than the number of words in the nodes themselves However, this turns out to give a degradation at the higher range Observing the result, the matches at the higher range have very few words in the child nodes If the number of child nodes are small in both ontologies and they have many in common, the distance between the nodes becomes extremely small This could be both beneficial and harmful It can pick up some matches which could not be found in Experiment-1, but the matches could be good or bad ones The following example is a good one which is actually found at the ninth rank in Experiment-4

```
EDR
 parent(*)  No English word, J-description
       "target animals hunting or fishing"
 children   game,kill
WordNet
 parent(*)  prey, quarry
 children   game
```

Adding parent nodes (Exp-5)

In Experiment-5, the words in the nodes themselves and their parent nodes are used It can be naturally thought that the words in the parent nodes are useful to disambiguate polysemous words The result confirmed this In the highest range, category A has 10 matches out of 20 which is three times as much as in Experiment-1, and twice that in Experiment-2 and 3

Using parents, self and children (Exp-6,7)

In Experiment-6 and 7 words in parent, self and child nodes are used with different weightings All evaluation results are identical except the lowest range, and these have the largest number of A's at the highest range among all of the experiments This indicates that three sources together is better than any two or any single source of information

Adding grandchild nodes (Exp-8)

Finally, in Experiment-8, words in all four kinds of nodes, parent, self, child and grandchild, are used The evaluation result is the same as that in

Experiment-6, and we could not see improvement by adding grandchild information Actually, by observing the result, we can see that the information at the grandchild level is not so useful

Observing the evaluation process

From the evaluation process, we understand that a human uses not only the four kinds of information, but also information in grandparent or the successor's nodes Some improvement might be obtained if we used such information Also, we might be able to achieve more improvement by using sibling nodes, and the result of distance calculation of other nodes

As we presented by the example of "mammal" and "elephant", there are the cases where in one ontology a relationship is parent-child, but in the other ontology it is a grandparent-grandchild relationship or a sibling-relationship It would be better if we took the characteristics of each ontology and differences of the ontologies into account in the calculation In particular, the information in ancestors might be very useful

Other distance definitions

In our method, we simply used the dice coefficient However, we can use more complicated or sophisticated measures For example, (Resnik 1995) proposed a measure of semantic similarity based on the notion of information content Although this proposal defines similarity between two nodes in a single taxonomy or ontology, we may be able to apply it in our situation

(Agirre et al 1995) proposed conceptual distance between nodes on ontologies captured by a Conceptual Density formula It is also a definition in a single ontology

Recently, (O'Hara and et al 1998) conducted an experiment of matching two ontologies, WordNet and the Mikrokosmos Ontology They used the definition proposed in (Resnik 1995) among other heuristics It is not so clear how to compare the method to our method, as they used several heuristics which is not directly comparable to our method However we noticed that it is a very important to investigate their methods

4 Conclusion

We proposed a statistical method of matching two ontologies Since it is impossible to exhaustively consider all matches by hand, automatic methods to make matches or to narrow down the candidate matches are needed Although the experiments are preliminary, they show what kinds of information is useful in statistical matching We found that parent nodes, besides the nodes themselves, are the most useful for matching by disambiguating the synonyms of words The best performance was achieved by using words in parent, itself and child nodes We observed that it is important to consider the characteristics of the ontologies One goal of our future work is to understand how to incorporate such characteristics into these statistical methods

5 Acknowledgements

We would like to thank Prof Ralph Grishman at New York University for his suggestions and anonymous reviewers who gave us some severe comments

References

Eneko Agirre and German Rigau, "A Proposal for Word sense Disambiguation using Conceptional distance" *Proc of the 1st International Conference on Recent Advances in natural Language Processing* 1995

EDR Electronic Dictionary Version 1 5 Technical Guide *EDR TR2-007*, 1996

Eduard Hovy "Creating a Large Ontology", *ANSI Ad Hoc Group on Ontology*, Stanford University, 1996

George Miller "WordNet A lexical database for English" *Communications of the ACM, 38(11) pp39-41*, 1995

Hideo Miyoshi, Kenji Sugiyama, Masahiro Kobayashi and Takano Ogino "An Overview of the EDR Electronic Dictionary and the Current Status of Its Utilization", *Proc of COLING-96*, 1996

Takano Ogino, Hideo Miyoshi, Masahiro Kobayashi, Fumihito Nishino and Jun'ichi Tsujii "An Experiment on Matching EDR Concept Classification Dictionary with WordNet", *Proc of IJICAI-97*, 1997

Tom O'Hara, Kavi Mahesh and Sergei Nirenburg, "Lexical Acquisition with WordNet and the Mikrokosmos Ontology" *Proc of the COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems* 1998

Pangloss Project (Information Sciences Institute (ISI) / University of Southern California (USC)) homepage "http //www isi edu/natural-language/nlp-at-isi html"

Philip Resnik, "Using Information Content to Evaluate Semantic Similarity in a Taxonomy" *Proc of the 14th International Joint Conference on Artificial Intelligence*, 1995

UTIYAMA Masao, HASHIDA Koiti, "Bottom-up Alignment of Ontologies" *Proc of IJCAI97 Workshop on Ontologies and Multilingual Natural Language Processing* 1997

WordNet Homepage "http //www cogsci princeton edu/ wn/"