Abstraction of the EDR Concept Classification and its Effectiveness in Word Sense Disambiguation

KIMURA Kazuhiro and HIRAKAWA Hideki

Human Interface Laboratory
Corporate Research & Development Center
TOSHIBA
1, Komukai-Toshiba-cho, Saiwai-ku, KAWASAKI 212-8582 JAPAN
{kazu.kimura,hideki.hirakawa}@toshiba.co.jp

Abstract

The relation between the degree of abstraction of a concept and the explanation capability (validity and coverage) of conceptual description which is the constraint held between concepts is clarified experimentally by performing the operation called concept abstraction. This is the procedure that chooses a set certain of lower level concepts in a concept hierarchy and maps the set to one or more upper level (abstract) concepts. We took the three abstraction techniques of a flat depth, a flat size, and a flat probability method for the degree of abstraction. By taking these methods and degrees as a parameter, we applied the concept abstraction to the EDR Concept Classifications and performed word sense disambiguation test. The test set and the disambiguation knowledge were extracted as a co-occurrence expression from the EDR Corpora. Through the test, we found that the flat probability method gives the best result. We also carried out an evaluation by comparing the abstracted hierarchy with that of human introspection and found the flat size method gives the most similar results to human. These results would contribute to clarify the appropriate detailed-ness of a concept when given an application purpose of a concept hierarchy.

Introduction

In Word Sense Disambiguation (hereafter WSD) problem, first we should make clear how to define the unit of a meaning. Moreover, it is an important subject to clarify the unit whether having sufficient description power, that is semantic discrimination ability, in respect to a certain application purpose. For example, in the Japanese to English machine translation system ALT-J/E (Ikehara et al. 1991), they use about 3,000 semantic categories with hierarchical construction. This classification, also known as the NTT-thesaurus (Ikehara el al. Eds. 1999), is said to give an enough discrimination ability to make a correct choice of target words. In WordNet (Fellbaum Eds. 1998) about 90,000 synsets are defined for English. And in the EDR Concept Classification, which was designed as an intermediate language for both English and Japanese, about 400,000 semantic nodes are classified. Of course, if the concept is defined more detailed, the discrimination ability will increase a certain degree. However, a problem arises when the concepts are defined in more detail, the cost of developing actual semantic constraints among a number of concepts. Furthermore, too detailed definition may lead to exceed the human ability of sense discrimination. Then the constraint development itself may be not reliable work as opposed to the expectations of concept designers. Hence we believe there exists a certain appropriate level of detailed-ness of a concept if given a certain application purpose, such as target word selection (or WSD) in machine translation and similarity measuring in information retrieval.

In order to explore about this problem, we carry out the operation called concept abstraction. This is the procedure that chooses a certain set of lower level concepts in a concept hierarchy and maps the set to one or more upper level (abstract) concepts. It clarifies the relation between the degree of abstraction of a concept and the explanation capability (validity and coverage) of conceptual description which is the constraint held between concepts. As an initial value of classifications and constraints, we adopt the Concept Classification and the Corpora provided by EDR because its classification is highly detailed than the other semantic systems and the corpora is semantically tagged.

In the remainder of this paper we will first of all outline a concept abstraction and then the three mechanical abstraction techniques of a flat depth, a flat size, and a flat probability method. By taking these methods and degrees of abstraction as a parameter we will describe the WSD test to clarify its precision and coverage. We will also make an evaluation by comparing the abstracted hierarchy with that of human introspection and then make a comparison between the abstraction techniques. And finally we will consider about the appropriate detailed-ness of a concept for WSD task.

Concept abstraction

Advantages of concept abstraction

Concept abstraction is a mechanical operation that simplifies a concept hierarchy. First, the terms used in this paper are enumerated.

Concept: A semantic label that is assigned one or more to one word. Hereafter, a hexadecimal number called concept identifier or parenthesis expression such as <*word*> is used to express a concept.

Concept Classification: A set of constraints that define an upper/lower relation between two concepts. It is also called *isa* relation, *ako* relation, etc. It constitutes a tree or a directed acyclic graph (DAG.) The latter is taken when a multiple inheritance is allowed. The set of extensions of a concept is a union of that of dominants. An intension of a concept, that is the conceptual description about the concept, also holds at its dominants. This is known as the inheritance mechanism.

Concept Description: A set of constraints that define a semantic relation among concepts. In this paper we only consider the constraints between two concepts, not more. As a semantic relation, we use a normal deep-case label such as agent, *object*, *scene*, etc. For notational convenience a 3-tuple of <*relation, concept1, concept2*> is used for a concept description.

Concept Abstraction: A procedure which chooses a certain concept set out of a given concept classification, and maps the set to one or more upper level concepts.

Abstraction of a concept is, for example, a mapping a concept <*thoroughbred*> to concept <*horse*> or concept <*animal*>. The number of conceptual nodes classified decreases and the load of system maintenance is mitigated by abstraction. Appropriate abstraction contributes to decrease unnecessary subdivision and provides the detailed-ness suitable for development of conceptual description. The detailed-ness may be expected to be flat that is a desirable feature especially for measuring semantic similarity. Furthermore, concept abstraction generalizes concept descriptions and spreads their coverage. For example, if a concept description <*agent, thoroughbreds, run*> is abstracted like <*agent, horses, run*>, then the original one now can be applied to *horses* other than *thoroughbreds*. Thus, the advantages of abstraction are summarized below.

To flatten concepts: Maintenance of a concept classification is needed constantly. Therefore, it is desirable to build it with the concepts of necessary minimum as possible, and that viewpoint of a classification is clear. Furthermore, it is expected to provide semantic discrimination ability suitable for the application purposes such as IR and MT, and to give a semantic unit that enables exact and compact concept description development. Concept abstraction contributes to these characteristics.

To generalize concept descriptions: A concept description is an approximation of general knowledge, and is mainly gradually gained from many examples (corpora). Concept abstraction generalizes concept description obtained from individual examples on an appropriate level, and expands the knowledge that may be applied for other examples. However, it is necessary to regard that excessive abstraction, say over-generalization leads a wrong knowledge, and lowers its reliability.

Techniques of concept abstraction

Here we introduce the three techniques of concept abstraction, which we implement and evaluate.

Flat Depth Method: The concept whose depth from the root is more than constant *D* is abstracted to its ancestor concept of depth *D*. If the hierarchy has the feature that every semantic distance (or similarity) between mother and daughter node is guaranteed to be fixed when the depth *D* is given, it will be thought that this technique is effective. However, it is difficult to assume the existence of such classification.

Flat Size Method: The descendants are abstracted to their ancestor so that the number of dominants of the ancestor to be in the neighborhood of a constant *S*. If each detailed-ness of given concept is almost same, the detailed-ness of abstracted one will be flat, and be considered that this technique is effective.

Flat Probability Method: The descendants are abstracted to their ancestor so that the appearance probability of dominants of the ancestor to be in the neighborhood of a constant *P*. The probability of the ancestor is calculated by adding sum of the probabilities of its dominants to the own probability. Then the probability of the root node equals one. This is the same approach by Li and Abe (1995), Hirakawa et al. (1996), and McCarthy (1997b). In contrast, there is another framework by Resnik (1993.) It ensures that the sum of probabilities over the entire hierarchy equals one. We didn't take this since flatness of leaf nodes is the matter of concern. Appropriate abstraction is expected if we can assume the appearance probability is proportional to the detailed-ness of a concept. However, when obtaining for appearance probability from a corpus, although it is hard to consider another choice, it should be an appropriate sample of the universe, such as all documents of a task domain to which concept descriptions are developed.

Treatment of a multiple inheritance in concept abstraction

A concept classification constitutes a tree or a DAG. The latter allows existence of two or more upper level concepts to a certain concept, and it inherits the attribute of these upper level concepts (multiple inheritance). For example, a Japanese word *uma* (a horse) has several correspondences to the concept label, such as a horse as an animal, a horse as edible meat, a horse as a vehicle, and a horse as a man of Japanese chess. If these are all considered to be polysemy, although a hierarchy will serve as the simple tree structure without a multiple

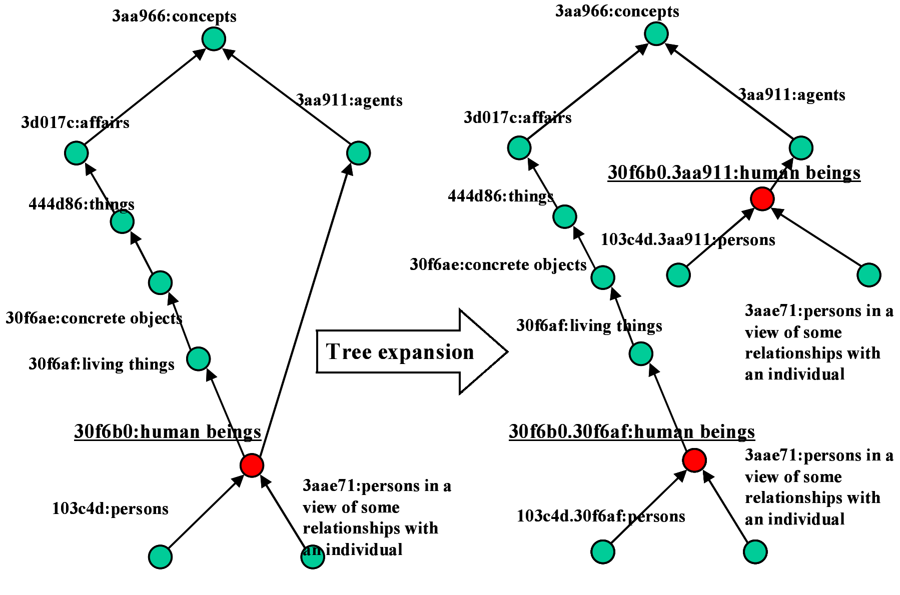

Figure 1: Multiple Inheritances in Concept Abstraction

inheritance, the number of concepts increases and concept description will become detailed. On the other hand, a whole or part of these polysemous labels can be designed to be one same conceptual label <*uma*> and it to be a descendant of two or more concepts, such as <*animal*> and <*edible meat*>. Then the hierarchy will take a DAG structure.

In general, it is a very difficult work to set up polysemy mutually disjoint. The development of concept descriptions requires the judgment of whether a statement carries the concept 1 or 2 in a tree. Since these are not guaranteed disjoint, they may be troubled in the judgment. However, since a DAG structure may allow suspending such judgment, it can be considered to be efficient structure in developing conceptual descriptions. Hence it is more general to take the DAG structure of accepting a multiple inheritance, in a practical concept classification.

Now, when considering concept abstraction, it arises the problem how the concept with this multiple inheritance should be abstracted. Figure 1 shows the view in this paper. A fundamental view is expanding the DAG to a tree and regarding a multiply inherited node to be different by the path from the root concept. For example, the concept "306b0:human beings" has two parents, one is the "30f6af:living things", and the other is the "3aa911:agents." When abstracting, there are two ways; one is abstracting to either of them. The other is abstracting to both of them. However, the former has the problem of throwing away one of attributes. The latter has the problem of producing redundant semantic ambiguity, since it produces one-to-many correspondences. The concept description should be abstracted to the proper one of them by nature. But in this paper, the latter is taken because the former problem is more serious. Therefore, for example, when the concept "103c4d:persons" is abstracted by the flat depth method of depth 2, it will be abstracted into both the "30f6b0:human beings" and the "444d86:things."

Concept abstraction of the EDR concept classification

EDR concept classification

This section describes the concept abstraction of the EDR Concept Classification.

When viewing a concept hierarchy widely, the so-called word thesaurus can also be in this category. For Japanese, there are several well-known thesauri other than the EDR, such as the one by the National Japanese Language Research Institute, the one by the Kadokawa publisher, and the NTT-thesaurus. Besides the NTT-thesaurus, many of them are not designed for computation and the vocabulary is rather small. The NTT-thesaurus as electronic data is available to limited researchers in academic site and the 3,000 nodes classification is rather small to abstraction purposes. Therefore, we took the EDR Concept Classification as the base system, which consists of 400,000 semantic labels.

The EDR Concept Classification and the other EDR Dictionaries have been developed by the ten years effort with the national project of MITI. The target languages are Japanese and English. The EDR Concept Classification is mutually associated through the hexadecimal number called concept identifier with other EDR dictionaries such as a word dictionary, a bilingual dictionary, and a corpus. The Word Dictionary gives correspondence between lexical information including word stem and grammatical features, and the conceptual information implemented as the concept identifier. There are about 200,000 words registered for each language. The Concept Classification does not have language dependency, which is the distinction between Japanese and English. The concept hierarchy consists of one root node, and intermediate nodes of 11,480, and leaf nodes of 376,432.

EDR concept classification admits the multiple inheritances stated above. More than 20,000 concepts among 400,000 have multiple inheritances. Since there are quite many of them, the treatment of their abstraction stated above should have a considerable effect on the experiment and evaluation about which we will argue from now on.

Concept abstraction algorithm

Before the concept abstraction, **Japanization of the Classification** should be carried out for the EDR Classification. This is because the concepts from the language of not in attention (in this case English) will cause side effects especially in the flat size and the flat probability method. Japanization of the classification is embodied as removing a concept identifier who relates only to an English Word dictionary. Although the EDR concept classification is language independent in design, almost portions of the constituents are language dependent in fact. The nodes common to both Japanese and English are 33,000. This seems to reflect the history that Japanese and English classifications developed individually. Therefore, it is quite natural to divide a concept hierarchy into Japanese and English. In addition, unclassified concepts were also deleted.

The algorithm of Japanizing classification is shown below.

Step 1. A leaf node set of the English origin (and the unclassified concept) is stored in a set *Delete*.

Step 2. A set of the mother nodes of the direct above of all the elements of *Delete* is stored in a set *Next*.

Step 3. Steps 4-5 are repeated until *Next* becomes an empty set.

Step 4. A set *Current* is stored all the elements of the *Next*. Then the *Next* is assigned an empty set.

Step 5. For each element of the *Current*, if all the direct daughters of it are included in the *Delete*, then the element is added to *Delete* and the set of the direct mother nodes of it is stored to the *Next*

Step 6. All the elements of the set *Delete* and the links related to them are deleted from the concept hierarchy.

By this algorithm, all the leaf and intermediate nodes that were only related to English are deleted and the rest of them turn into the Japanese concept hierarchy. Thus we got the hierarchy of 199,245 conceptual nodes.

Next, the algorithm of the flat depth method is described. By the treatment of a multiple inheritance described above, that is the tree expansion, the following simple algorithm is derived.

Step 1. While traversing the hierarchy in depth first, if the node *n* whose depth is less than a specified value is encountered, then the node *n* and all the descendants of it are abstracted to the direct mother node of *n*.

The result of abstraction is saved as a correspondence table of the original concept identifier and those of after abstraction Plurality of abstraction is allowed. This table is henceforth called **abstraction map**. The maps were created by changing the parameters from depth=1 to 15.

Next, the algorithm of the flat size method is described.

Step 1. For each node of the hierarchy, the number of dominated nodes is computed and stored on DB. The node with a multiple inheritance is double counted since from the view of the tree expansion.

Step 2. While traversing the hierarchy in depth first, if the node *n* whose number of dominants is less than a specified value is encountered, then the node *n* and all the descendants of it are abstracted to the direct mother node of *n*. When *isa* relation holds between the abstracted nodes of a node, only the node of the minimum distance is adopted[1] (**minimum abstraction**).

The abstraction map was created for 15 parameters from size=1 to 50000 by this algorithm.

About the flat size algorithm, although the method by Hearst and Schütze (1993) is known, we didn't take the one since the original graph structure will be broken. It is not a critical problem for them since it aims at word similarity calculation.

Next, the algorithm of the flat probability method is described. This is the almost same algorithm as the flat size method.

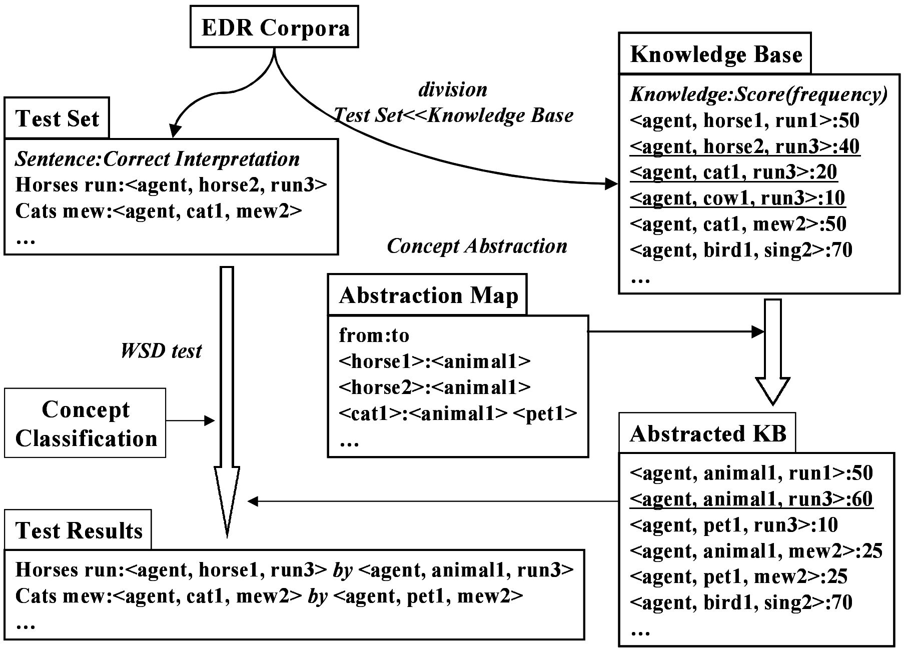

Figure 2: Evaluation by WSD test

Step 1. The frequency of the concept of a content word that appeared in the EDR Corpora is counted.

Step 2. For treating sparseness problem, Good-Turing Discounting (Good 1995) is carried out by following the equation (1).

$$r^* = \frac{(r+1)n_{r+1}}{n_r} \qquad (1)$$

where *r* is an observed frequency, n_r is a number of the kind of the concept that appeared just *r* times, and *r** is the discounted frequency of *r*.

It discounts a certain amount of observed frequency, that is the sum of *r*-*r**, and distributes it equally to the non-observed concepts. Thus non-observed concepts were assigned frequency 0.16, probability 0.00000009.

Step 3. For each node of the hierarchy, the frequency with above discounting is computed and stored on DB. The frequency of a node is added the sum of all frequencies of its dominants. The frequency of a node with a multiple inheritance is divided equally and summed up to its mothers.

Step 4. While traversing the hierarchy in depth first, if the node *n* whose frequency is less than a specified value is encountered, then the node *n* and all the descendants of it are abstracted to the direct mother node of *n*. Furthermore, the minimum abstraction is performed.

By this algorithm, the abstraction map was created for 17 parameters from probability=0.00001 to 0.2.

Evaluation by the WSD task

WSD task

As for quantitative evaluation of concept abstraction, the WSD task using the EDR Corpora is described. The outline of the task is shown in Figure 2. The EDR Corpora provides not only a syntactic tag but also a semantic tag in a form of concept identifier. Hence it gives a number of semantic co-occurrence relations. For the WSD experiment, we use these co-occurrence data instead of a whole sentence for simplicity. These co-occurrence data can be easily extracted from the corpora. Then they are divided into two portions, a test set and a knowledge base.

The test set entry is a 2-tuple of a sentence and its correct interpretation. Here the sentence is a simple noun-verb co-occurrence expression and the interpretation is a 3-tuple of the corresponding concept identifiers and their relation expressed by a deep case label. The knowledge base entry is the same form as the interpretation and is also attached its frequency in the corpora.

Thus we got 14,000 of test set entries and 500,000 of knowledge base entries. The semantic ambiguity, that is the number of assigned concept identifiers to word 1 *

[1] On the other hand, we can consider the **maximum abstraction** that adopts the node of the maximum distance. We didn't take this since there is a case where the relation (or axiom) held before abstraction no longer holds after abstraction. In addition, it should be noted that the flat depth method does not perform the minimum abstraction but treat all as ambiguity. This is because the map to both itself and the mother is possible, and if the minimum abstraction is performed at this time, there is a case where the knowledge (or concept description) that should be abstracted from the viewpoint of depth is not abstracted.

that to word 2, was 16.4 in average. The frequency associated with the knowledge base entry is considered to represent the reliability of it and is used as score to select the best interpretation from possible ones. The score is given by the following formula in consideration of the position in the hierarchy and frequency.

$$score(<rel,c1,c2>) = \frac{freq(<rel,c1,c2>)}{dom(c1)\cdot dom(c2)} \quad (2)$$

where *freq(<rel,c1,c2>)* is the observed frequency of concept description *<rel, c1, c2>* and *dom(c)* is the number of dominated nodes of concept *c*.

Since the relation between some intermediate node *c1* and *c2* can be thought as the abstraction of the relations between all the leaves of c1 and c2, this equation works to give an equivalent score which all the leaf-leaf relations are given to.

Next the algorithm for the WSD test is described. The test is performed for each abstraction parameter of the flat depth, size, and probability, and for the baseline mapping that every concept is mapped into itself.

Step 1. A knowledge base is abstracted using the abstraction map. That is, for each concept description, each concept identifier in the description is mapped into the abstract one by the abstraction map while the relation remains the same. The score of the abstract concept description is the sum of that of the original concept description. Since the mapping may not be unique as described above, the abstraction of concept description may have *n* possibilities. In this case, *s/n* is given to the score of abstract concept description provided the original score *s*. In the Figure 2 example, *<agent, cat1, run3>* is abstracted into *<agent, animal1, run3>* and the score from the original one will be given 20/2 since *<cat1>* has two possibilities of abstraction.

Step 2. For each possible interpretation *<rel, c1, c2>*of a test sentence, we select the interpretations that are supported by the abstracted knowledge base and the inheritance on the hierarchy, and order them by the score. When at least one interpretation is selected, we say the WSD was done. The number of WSD performed is counted by the counter *c*.

Step 3. When the correct interpretation exists among the top-ranked *N* interpretations, we say the WSD is precise and we give *1/N* count to the counter *pj*.

Step 4. Finally we calculate the precision *P*, the coverage *C*, and the F-measure *F* by following equations.

$$P = \frac{\sum_j p_j}{c} \quad (3)$$

$$C = \frac{c}{N} \quad (4)$$

$$F = \frac{(\beta^2+1)C\cdot P}{\beta^2\cdot P + C} \quad (5)$$

where *N* is the total number of the test sentences and *β* is the coefficient that weighs the relative importance of *C* to *P*. *C* and *P* are equally important when *β*=1, and we will show the result by that value.

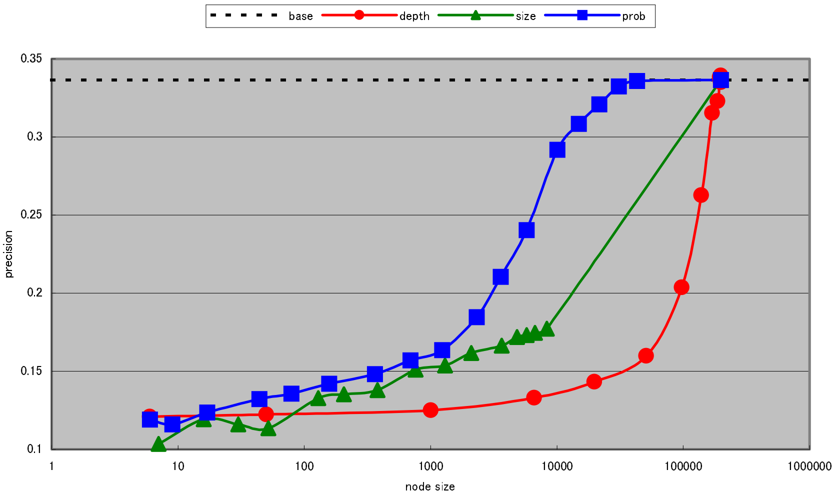

Figure 3: Precision

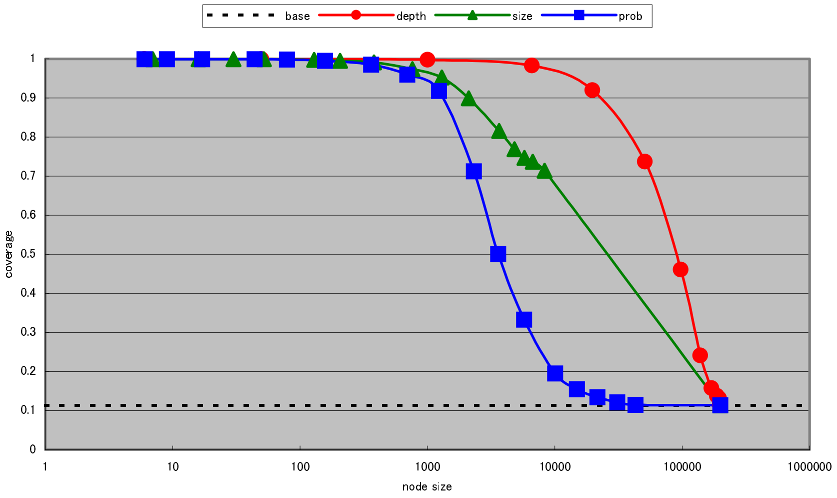

Figure 4: Coverage

Evaluation results

The results of the evaluation are shown in Figures 3, 4, and 5. In order to see the difference among the abstraction techniques, the unit of horizontal axis takes the total number of the abstracted nodes instead of the flatting parameters. In addition, the dotted line shows the result of the baseline evaluation.

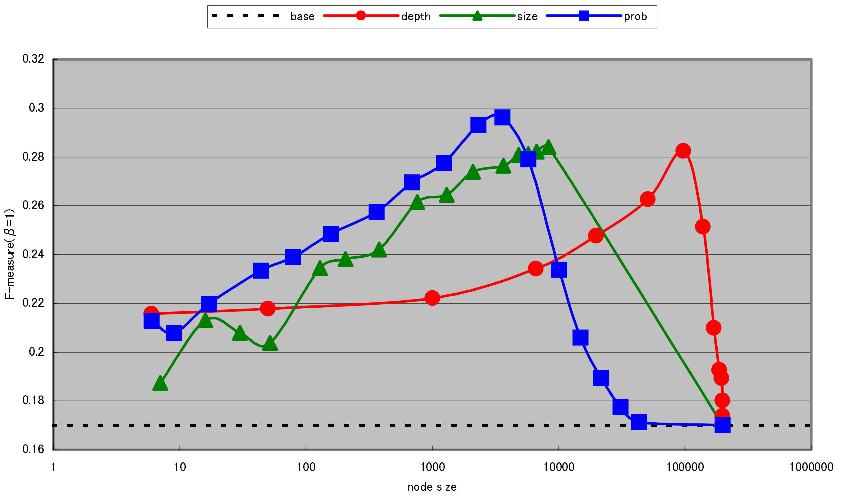

Figure 5: F-measure

Comparison with human introspection

In addition to the WSD evaluation, we also make a comparison with the abstraction by human introspection. The abstraction data are built by sampling 100 EDR leaf nodes and abstracting them. In abstracting we take two ways. One is the use of the NTT-thesaurus and the other is the use of our own intuitions. The NTT-thesaurus is designed to make a correct target word selection in machine translation. Therefore it can be a standard classification to provide the detailed-ness of a concept appropriate for WSD. To make the comparison, for all the sampled EDR nodes, we observe the path from the node to the root and choose one intermediate node that is considered to correspond to an NTT's leaf node. This results the abstraction map for 100 sample nodes.

In the same way, by using our own intuitions, we get another abstraction map. In choosing abstracted node, two intuitive criteria are used. One is that the concept should be able to express by one typical word, neither compounds nor phrases. The other is that the concept should be stereotyped that one can concretely imagine its typical shape, character, etc. For example, the concept <*skirts*> may be chosen but <*wears*> may not because it is too abstract and variant to imagine its own shape. Although we don't have psychological evidence to take this criterion, we guess these levels of concept will correspond to the inter-lingual Common Base Concepts in the EuroWordNet.

After these two abstraction maps are created, each is compared with that of the mechanical methods described before. The measure of comparison is a distance in mean between the abstracted nodes by introspection and those by a computer. Here the distance is the number of edges in the shortest path of the two nodes.

The results are shown in Figure 6 and 7.

Discussion

Precision

The baseline precision is 33.6% although the better rate was expected. The main reason would be the sparseness problem that the WSD knowledge is still insufficient. The form of the graph was as expected. Since the abstraction is an operation that makes an expanded interpretation of the observed knowledge, it basically decreases according as the degree of abstraction is high or the total abstracted nodes is small. However, by the flat depth method, the result exceeds the baseline a little among the depth11-14. In this interval, the total number of abstracted nodes does not change with the baseline so much and the abstracted knowledge is also keeping the low degree of abstraction. Hence the abstraction seems to work only in the direction of compensating the sparseness of knowledge. As for the comparison among three techniques, it is converging on the baseline with the small number of nodes in the order of the flat probability, the flat size, and the flat depth method. The flat probability method provides the almost same discrimination ability of meaning as the baseline with the fewest number of nodes.

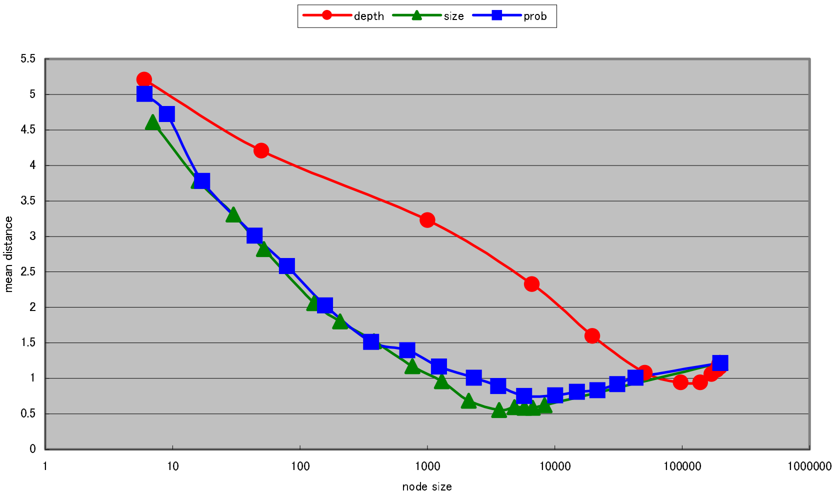

Figure 6: Comparison with the NTT-thesaurus

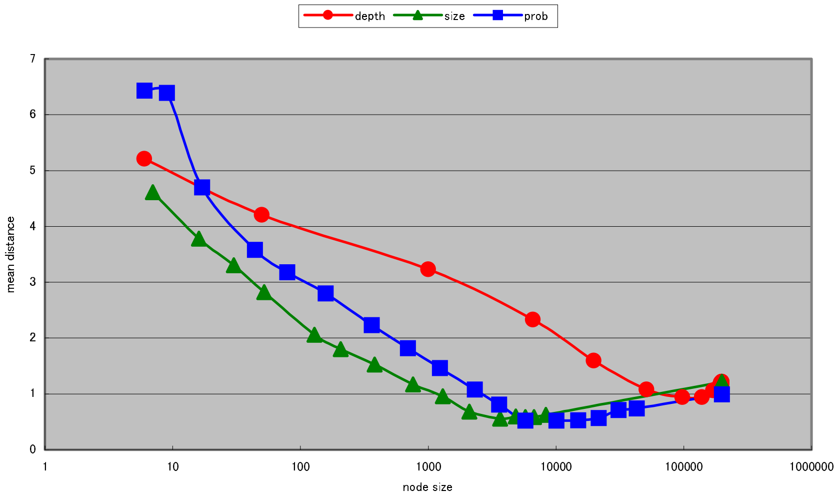

Figure 7: Comparison with the abstraction by our own intuition

This was a result as expected.

Coverage

The baseline coverage is 11.4%. It is obvious that the coverage increases when the abstraction level becomes higher. The coverage is converging on 100 % with the small number of nodes in the order of the flat depth, the flat size, and the flat probability method. Of course, the evaluation by coverage alone is meaningless since overgeneralization leads wrong knowledge. Therefore, we should see the integrative evaluation by F-measure to judge the superiority among the abstraction techniques.

F-measure

A baseline F-measure is 17.0%. Each abstraction technique basically works to raise the coverage and to lower the precision. When seeing integrative, all the result exceeds the baseline although it depends on how to take the value β. (Note that β=0 results the precision.) As for the comparison among three techniques, the local maximum has appeared with the few number of nodes in the order of the flat probability, the flat size, and the flat depth method, and the flat probability method gives the best result again. This tendency holds when weighing the precision more by β=0.5

The local maximum of the flat probability method is 29.6% at the point of 3,583 nodes of abstraction level. In our research community it has been said empirically that the appropriate number of nodes, which constitute a concept hierarchy, will be about 3,000. Our result agrees to this value.

In addition, in the flat probability method we also created another abstraction map without the Good-Turing discount, which is the direct use of observation frequency. But in all the evaluations of precision, coverage, and F-measure, it did not result a significant difference. We suspect that the existence of knowledge is rather important than its frequency in this experiment, since the training corpora from which the probability is estimated and the test set used for evaluation are in the same fields. They mainly composed of newspaper articles and magazine reports.

Comparison with the NTT-thesaurus

Figure 6 shows that the flat size and the flat probability method result a strong resemblance but the flat depth does not. This suggests a structural resemblance between the EDR and the NTT hierarchy. In both of them, the location of the leaf node varies in depth. Since the flat depth method does not reserve this property but the others does, its error will become more significant when comparing in the same node size. The abstraction that gives the most resemble hierarchy to the NTT's one, that is the local minimum of the graph, is the mean distance of 0.55 with its variance of 0.43, at the point of 3,650 nodes of abstraction level. Note that this level agrees the best result of the WSD test, although it may be made by chance.

In comparing the flat size and the flat probability method, the latter shows the best result at the less abstract level. That is the point of 5,743 of abstraction level and the mean distance of 0.75 with its variance of 0.57. The reason will be that the node that was given a probability keeps the distance from the NTT level since it remains un-abstracted until it reaches to a certain abstraction level

Comparison with the intuitive abstraction

Figure 7 shows the same tendency as above that the result by the flat size and probability method resembles although their difference became more salient. The flat size method results the best again, at the point of 8,332 of abstraction level and the mean distance of 0.41 with its variance of 0.31. The abstraction level was lowered in comparison with the NTT case. This comes from the criteria for the intuitive abstraction. The nodes chosen for the test were more detailed than that for NTT case. It would also reflect that our perceptual unit of basic meaning is more detailed than the unit required for WSD.

Conclusion

This paper proposed the mechanical abstraction techniques to change the detailed-ness of a concept hierarchy. By using these techniques, we considered about the appropriate detailed-ness of a concept when given an application purpose of a concept hierarchy.

Through the WSD experiment, we clarified that the concept abstraction contributes WSD more than using a hierarchy as it is, and the flat probability method gives the best result among the three techniques proposed here.

We also carried out an evaluation by comparing the abstracted hierarchy with that of human introspection. In this case the flat size method gives the most similar results to human.

In each case, the 3,600 nodes level of abstraction is considered to be appropriate for WSD, especially the WSD for machine translation. For information retrieval, that is another useful application, it was not clarified in this paper, although the comparison with the intuitive abstraction by human introspection suggests a perceptual unit of basic meaning, which would have a strong relevance to word similarity measure.

For future work, we will first reconsider the treatment of probability within the framework of statistics. We used frequency comparatively direct without considering its statistical reliability. This will make the experiment more precise, although the overall tendency will not change.

Secondly, we will enrich the data of human introspection. This time just one rater was engaged in the work. By using several raters' data, the reliability of the standard data will be measurable by inter-rater reliability. The use of reliable date will make our results more precise.

Finally, the approach to the sparseness problem is still an important issue. Semantically tagged corpora were available in this time by employing the EDR's one, but we cannot increase the amount anymore. There are a lot of researches into automatic semantic taggers, especially by making a use of word frequency in raw corpora and a hierarchically constructed word thesaurus. We are now planning to take these advantages of them.

References

EDR (Eds.) (1996). EDR Electronic Dictionary Version 1.5 Technical Guide. EDR Technical Report TR2-007.

Fellbaum, C. (Eds.) (1998). WORDNET -- An Electronic Lexical Database. The MIT Press.

Good, I. J. (1953). The Population Frequencies of Species and the Estimation of Population Parameters. In Biometrika, 40, pp.237-264.

Hearst, M. and Schütze, H. (1993). Customizing a Lexicon to Better Suit a Computational Task. In Proceedings of the ACL SIGLEX Workshop.

Hirakawa, H., Xu, Z., and Haase, K. (1996). Inherited Feature-based Similarity Measure on Large Semantic Hierarchy and Large Text Corpus. In Proceedings of the COLING-96, pp.508-513.

Ikehara, S., Shirai, S., Yokoo, A., and Nakaiwa, H. (1991). Toward an MT System without Pre-Editing -- Effects of New Methods in ALT-J/E --. In Proceedings of MT Summit III, pp.101-106.

Ikehara, S. et al. (Eds.) (1999). Nihongo-Goi-Taikei CD-ROM Version. Iwanami Shoten. (in Japanese.)

Li, H. and Abe, N.(1995). Generalizing Case Frames Using a Thesaurus and the MDL Principle. In Proceedings of the International Conference on Recent Advances in Natural Language Processing, pp. 239-248 Bulgaria.

McCarthy, D. (1997a). Word Sense Disambiguation for Acquisition of Selectional Preferences. In Proceesings of the ACL/EACL 97 Workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications, Madrid Spain, pp. 52-61.

McCarthy, D. (1997b). Estimation of a Probability Distribution over a Hierarchical Classification. In the Tenth White House Papers COGS – CSRP.

Resnik, P. (1993). Selection and Information: A Class-Based Approach to Lexical Relationships. Ph.D. thesis, University of Pennsylvania.

Resnik, P. (1995a). Disambiguating Noun Groupings with Respect to WordNet Senses. In Proceedings of the ACL-95

Resnik, P. (1995b). Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In Proceedings of the IJCAI-95.