Bottom-up Alignment of Ontologies

UTIYAMA Masao
Shinshu University
500 Wakasato, Nagano-si,
Nagano 380 Japan

HASIDA Kôiti
Electrotechnical Laboratory
1-1-4, Umezono, Tukuba,
Ibaraki 305 Japan

Abstract

A large scale multilingual ontology is necessary in order to support multilingual communication across the Internet by AI applications such as machine translation and multilingual information retrieval. To build such an ontology, alignment of EDR Concept Classification Dictionary and WordNet has been attempted. An experiment shows that the two ontologies are very different with respect to the distances among lexical entries. Another experiment on alignment of them based on a maximum weight matching algorithm shows that more than a half of the correspondences can be automatically recognized even in highly ambiguous cases.

1 Introduction

The explosive spread of the Internet is fermenting a huge bunch of desire of people to communicate their ideas to each other. However, this desire is yet far from fully satisfied. One major reason for this is language barrier. For example, if your web pages are written in Japanese, it is unlikely that they acquire a huge readership around the globe. Another major reason in this connection is the lack of computational infrastructure to support translation, retrieval, extraction, and so on.

The GDA (Global Document Annotation) initiative attempts to solve these problems by:

- announcing a common SGML (or XML) tagset to annotate electronic documents in such a way that machines could understand their semantic and pragmatic structures,
- providing tag-based AI applications (such as machine translation, information retrieval, and so on), which aid (multilingual) communication via electronic documents, and
- thereby having people annotate their electronic documents (mainly web pages) with the standard tagset.

Web authors will be motivated to annotate their pages if the annotated pages are translated, retrieved, and so on with a high accuracy and thus have greater chance to reach many, right kind of readers. As a result, a huge amount of annotated data is expected to emerge, which will trigger further supply of tag-based AI applications. Thus a positive feedback cycle will start turning. Namely, when plenty of high-quality, low-cost tag-based AI applications are available for communication aid, many authors will be motivated to markup their web pages, and when plenty of tagged documents are out there, more competitive tag-based AI applications will be provided, and so on and so forth.

Figure 1 should give a flavor of the tags to be proposed in GDA. A <seg> ⋯ </seg> encodes parse tree bracketing, and the property sem disambiguates polysemy of words. Note that these tags reduce the notorious ambiguities involved here, so that it is easy to automatically determine the underlying structure of the sentence by the present technology. The tags will also encode coreferences, scopes of logical/modal operators, rhetorical structure, social relationship between the author and the readers, and so on, in order to render documents machine-understandable to various degrees. The GDA tagset will be designed by incorporating the results of TEI[1], EAGLES[2], CES[3], and so on, as much as possible. The GDA tagset and a tagging editor (a tool to support tagging) are currently under development. They will be

```
<seg sem=time0>time</seg>
<seg>
  <seg sem=fly1>flies</seg>
  <seg sem=like0>like</seg> an arrow
</seg>
```

Figure 1: Annotated Text

[1] http://www.uic.edu:80/orgs/tei/

[2] http://www.ilc.pi.cnr.it/EAGLES/home.html

[3] http://www.cs.vassar.edu/CES/

published in August 1997 at:

http://www.etl.go.jp/etl/nl/GDA

for public review, discussion and experiments.

GDA can be kicked off with a minimal tagset which captures just parse-tree bracketing, for instance. Even such a simple tagset will no doubt dramatically improve the accuracy of machine translation, information retrieval, and so on. In order to maximally improve the quality of such GDA applications, however, tags should capture more detailed linguistic information, among which the most important will be lexical semantics. Tags should be rich enough to capture the linguistic semantics of various web pages in various languages. A large multilingual ontology is thus necessary after an initial stage of GDA; It is desirable that the values of the sem attribute be based on such an ontology.

In this connection, note that GDA does not only need such an ontology, but also promotes its development. That is, if the early stages of GDA prove useful even without sense tags, the development of a huge, sharable ontology will be strongly motivated in order to further improve the quality of GDA applications. The maintenance of the ontology will also be facilitated through the massive use of it by GDA applications. It is a hope of the authors that GDA should provide concrete objectives for which to promote international cooperations to develop a common multilingual ontology. The rest of the paper reports on some attempts toward developing a multilingual ontology for GDA by connecting various existing ontologies.

A connection among different ontologies may be tight or loose. A tightly connected ontology is a unified ontology into which the source ontologies have been mapped so that concepts originating from different ontologies are either unified or related in the same way as those originating from the same ontology are; Typically, the fact that one concept is a subset of another is reflected in the ontology in such a way that the former is a descendant node of the latter in the concept hierarchy. On the other hand, a loosely connected ontology is one in which the source ontologies are more loosely aligned, the concept hierarchy failing to capture some set-theoretic relations among concepts from different ontologies. In such an ontology compiled from an English ontology and a Japanese ontology, "brother" in the English ontology might not be connected with "*ani*" (elder brother) in the Japanese ontology.

High-quality, multilingual, tag-based applications would be easy to make if a tightly connected multilingual ontology were available. However, the development of such an ontology will require a huge cost of manual work. On the other hand, a loosely connected ontology may be automatically compiled by machines with relatively little human intervention. So such an ontology is much less expensive to build.

GDA can smoothly move from a loosely connected ontology to a tightly connected one, because annotation using a loosely connected ontology will still be effective under a tightly connected ontology as long as the former ontology is included in the latter. So tag-based applications presupposing a loosely connected ontology can be adapted to a tightly coupled ontology with minimum modifications. Under these considerations, GDA attempts to first demonstrate the usefulness of tags even without ontologies, and then to develop a loosely connected ontology, showing how tag-based applications based that ontology can aid multilingual communication efficiently, which will induce the motivation to develop a tightly connected ontology for more useful applications.

As a first step to construct a loosely connected ontology, the authors have examined the quantitative correspondences between two existing ontologies — EDR Concept Classification Dictionary [Yokoi, 1996] and WordNet [Miller, 1995] — and studied how to align them. What follows is a progress report on that.

2 Two Ontologies

2.1 EDR Concept Classification Dictionary

EDR concept classification Dictionary (hereafter EDR) is a bilingual ontology which contains a classification of concepts organized with respect to super-sub (is-a) relation. A concept may refer to some English and/or Japanese words (and phrases)[4] and is associated with an English explanation and a Japanese explanation of its intuitive sense.

Although EDR is a bilingual ontology, the Japanese part is not considered in the present paper because it is not needed to align EDR and WordNet.

2.2 WordNet

WordNet is an English ontology. It organizes English words and phrases into synonym sets ("synsets"), each representing one underlying lexical concept. Synsets are linked with each other via relationships such as super-sub and antonym. Approximately one half of the synsets have short English explanations of their intuitive sense.

Concepts and synsets are equally regarded as sets of words (and phrases) in the rest of the paper.

3 Comparison between EDR and WordNet

Two experiments were conducted to compare distances between words contained in the two ontologies.

[4] Some concepts do not refer to any words or phrases at all.

The distance between two words in each ontology is defined as follows. First, the distance D_E between two words w_1 and w_2 in EDR is defined as

$$D_E(w_1, w_2) = \min_{\substack{c_1 \in C(w_1) \\ c_2 \in C(w_2)}} d_E(c_1, c_2)$$

where $C(w_i)(i = 1, 2)$ is the set of concepts which contain w_i and $d_E(c_1, c_2)$ is the length of the shortest path between two concepts c_1 and c_2 in EDR. Second, the distance D_W between two words w_1 and w_2 in WordNet is defined as

$$D_W(w_1, w_2) = \min_{\substack{s_1 \in S(w_1) \\ s_2 \in S(w_2)}} d_W(s_1, s_2)$$

where $S(w_i)(i = 1, 2)$ is the set of synsets which contain w_i and $d_W(s_1, s_2)$ is the length of the shortest path between two synsets s_1 and s_2 in WordNet.

3.1 Comparison of WordNet Distance with EDR Distance

The first experiment concerns the distances in WordNet between two words (of a common syntactic category) between which the distance in EDR is zero. The result is shown in Table 1. In the table, "Num. of pairs" is the number of pairs of words whose WordNet distances are D_W.

The table shows that the number of pairs decreases as D_W increases. However, the decrease is not sharp. This implies that words which are close to each other in EDR are not necessarily close to each other in WordNet.

3.2 Comparison of EDR Distance with WordNet Distance

The second experiment examines the EDR distances between two words (of a common syntactic category) whose WordNet distance is zero. The result is shown in Table 2. As before, "Num. of pairs" is the number of pairs of words whose WordNet distance is D_E.

The decrease in the number of pairs shown in Table 2 is sharper than that in Table 1. But the decrease is not very sharp, either, which implies that words which are close to each other in WordNet are not necessarily close to each other in EDR.[5]

3.3 Remarks

Suppose that two words are far from each other in an ontology whereas they are close to each other in another. This gap arises in one of the following two cases.

Case 1. The words are actually similar in meaning, but the former ontology fails to capture that similarity.

Case 2. The words are actually dissimilar, but the latter ontology incorrectly connects them via a short path.

Case 1 occurs when the builder of the former ontology has overlooked some relations between the words. On the other hand, Case 2 occurs when the builder of the latter ontology has assigned wrong relations between the words. It is likely that Case 1 occurs much more frequently than Case 2. In fact, this was the case in the above two experiments according to the first author's survey.

Alignment of ontologies will improve the accuracy of the resulting ontology in two ways. First, missing edges in one ontology may be supplied by another, corresponding to Case 1. Second, erroneous edges may be removed, corresponding to Case 1, though that would be harder than the first type of improvement.

4 Alignment of EDR and WordNet

If an EDR concept and a WordNet synset are equivalent, very probably they share a word. We can conceive of a bipartite graph G by regarding this relation of sharing a word as an edge connecting the two nodes (the EDR concept and the WordNet synset). G consists of four subgraphs, G_{noun}, G_{verb}, $G_{adjective}$, and G_{adverb}, which correspond to the four syntactic categories in WordNet. $G_{cat}(cat = noun, verb, adjective, adverb)$ is the bipartite graph $(V_E^{cat}, V_W^{cat}, E_{EW}^{cat})$ where $V_E^{cat}(V_W^{cat})$ is the set of EDR concepts (WordNet synsets) containing words categorized in *cat*, and E_{EW}^{cat} is the set of edges connecting EDR concepts in V_E^{cat} and WordNet synsets in V_W^{cat}. A small part of G_{noun} is shown in Figure 2. In the figure, each box in the left column represents an EDR concept and contains the identifier of the concept (c_1, c_2, or c_3) and the words contained in the concept. Similarly, each box in the right column represents a WordNet synset.

Two experiments are reported in this section, one examining G and the other attempting to align EDR and WordNet based on this examination.

4.1 Number of Connected Components

The first experiment examined sizes of connected components in G. The size of a connected component can be very large. In that case, an EDR concept may have many candidates for the corresponding WordNet synset so that deciding the corresponding synset correctly will be difficult, and vice versa. There are large connected components mainly because of polysemous words. In Figure 2, for example, the polysemous word, "lime", connects c_2 to the semantically unrelated synsets, s_1 and s_3, so that the connected component spreads.

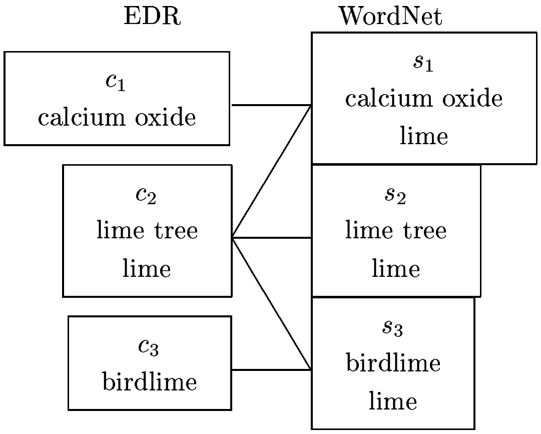
[5] In Table 2, the sudden decrease at $D_E = 1$ is caused by the structure of EDR. That is, $D_E = 2$ is more likely than $D_E = 1$ because the concepts which refer to some words are usually linked to the concepts which refer to no words in EDR.

D_W	0	1	2	3	4	5	6	7	8	9	over 10	total
Num. of pairs	11671	8993	9698	5943	5303	5640	5768	5613	5125	4448	21099	89301

Table 1: WordNet distance

D_E	0	1	2	3	4	5	6	7	8	9	over 10	total
Num. of pairs	11193	828	12828	3482	3966	3721	3741	4186	4007	3113	9962	61027

Table 2: EDR distance



This graph is part of G_{noun}. c_1, c_2, and c_3 are EDR concepts and s_1, s_2, and s_3 are WordNet synsets.

Figure 2: Bipartite-graph relation between EDR and WordNet

The distribution of connected components (CCs) in G is shown in Table 3. In the table, $|V_E|$ and $|V_W|$ stand for the number of EDR concepts and WordNet synsets, respectively, in each connected component. "$x \sim y$" means that the greater of $|V_E|$ and $|V_W|$ is in the range of x through y. The common syntactic category of the words in the connected component is shown in the table where "Num. of CCs" is one.

Table 3 shows that most connected components are small. Such connected components can be manually checked by humans to validate the edges therein. To examine the largest four connected components in Table 3, however, some computational aid by machine is necessary.

4.2 Alignment by Maximum Weight Matching Algorithm

The second experiment addresses an alignment of EDR concepts and WordNet synsets by extracting a one-to-one correspondence from G.

A one-to-one correspondence between EDR concepts and WordNet synsets is represented by a set of edges in which no two edges have common concepts or synsets. Such a set is called a *match* on G. In Figure 2, for example, $\{\langle c_2, s_1\rangle, \langle c_3, s_3\rangle\}$ is a match. We should probably generalize the notion of a match to allow one-to-many and many-to-one (but maybe not many-to-many) correspondences between concepts and synsets. The present formulation is a simplification for the sake of quick computation.

If each edge e in a match M has some real-number weight $w(e)$, then $w(M) = \sum_{e \in M} w(e)$ is called the weight of M. If $w(M)$ is the greatest of all the weights of matches in G, then M is called the *maximum weight match* (MWM) in G. Note that a match and a MWM can naturally be defined for subgraphs of G. For example, they can also be defined on the graph in Figure 2 which is a subgraph of G. Suppose that the edges in Figure 2 have an equal weight of one, then the MWM is $M' = \{\langle c_1, s_1\rangle, \langle c_2, s_2\rangle, \langle c_3, s_3\rangle\}$ and $w(M') = 3$.

If M is the MWM in G, then M will give the best one-to-one correspondence between EDR and WordNet on the condition that each edge in G is given an appropriate weight.

Let $e = \langle c, s\rangle$ be an edge where c is an EDR concept and s is a WordNet synset. Then the weight $w(e)$ of e is defined by

$$w(e) = \frac{|E \cap W|}{|E \cup W|} \sum_{x \in E \cap W} -\log P(x) \qquad (1)$$

where x is a word, and

$$\begin{aligned} E &= \{x | (x \in c' \vee x \text{ is in the explanation of } c') \\ &\qquad \wedge\ c' \text{ is a concept such that } d_E(c, c') \leq 1\} \\ W &= \{x | (x \in s' \vee x \text{ is in the explanation of } s') \\ &\qquad \wedge\ s' \text{ is a synset such that } d_W(s, s') \leq 1\} \\ P(x) &= \frac{\text{number of synsets containing } x}{\text{number of all synsets}}. \end{aligned}$$

See Section 3 for the definitions of d_E and d_W. In formula (1), $\frac{|E \cap W|}{|E \cup W|}$ captures the similarity between E and W. Due to $\sum -\log P(x)$, $w(e)$ takes a greater value when rarer words are shared by c and s.

The second experiment obtained the four MWMs listed in Table 4 for the four largest connected compo-

$\|V_E\|$	0	1	1	$2 \sim 5$	$6 \sim 10$	$11 \sim 15$	$16 \sim 33$	1531	11460	18841	34659
$\|V_W\|$	1	0	1					924	7345	7826	16969
Num. of CCs	22774	119095	15582	11908	652	92	40	1	1	1	1
category								adverb	adjective	verb	noun

Table 3: Number of connected components

label	category	V_E	V_W	size of MWM
A	adverb	1531	924	796
B	adjective	11460	7345	6416
C	verb	18841	7826	7608
D	noun	34659	16969	15948

Table 4: Connected components examined

nents in Table 3.

To examine the results, the edges in each MWM were first sorted in the decreasing order of the weights, and three parts (the top 30 edges, the middle 30 edges, and the last 30 edges) were extracted of each sorted list. Next, the correctness of the extracted edges were judged manually by the first author. The edges were classified into three categories: "correct" "wrong" and "uncertain."

The results of the judgments are shown in Table 5 through 8. In the tables, "range" means the range of the weights of the extracted edges. "av. degree" is the average number of edges connected with an EDR concept or a WordNet synset in a connected component.

edges	correct	wrong	uncertain	range
top 30	28	2	0	28.1 - 10.6
middle 30	17	6	7	2.6 - 2.4
bottom 30	10	14	6	0.4 - 0.0
total	55	22	13	28.1 - 0.0

av. degree = 3.74

Table 5: Judgment on edges extracted from A

edges	correct	wrong	uncertain	range
top 30	24	2	4	53.6 - 25.8
middle 30	12	14	4	3.1 - 3.0
bottom 30	6	15	9	0.3 - 0.1
total	42	31	17	53.6 - 0.1

av. degree = 5.52

Table 6: Judgment on edges extracted from B

edges	correct	wrong	uncertain	range
top 30	21	8	1	226.0 - 49.0
middle 30	14	9	7	4.3 - 4.2
bottom 30	10	14	6	0.4 - 0.0
total	45	31	14	226.0 - 0.0

av. degree = 8.47

Table 7: Judgment on edges extracted from C

edges	correct	wrong	uncertain	range
top 30	24	5	1	1093.01 - 102.91
middle 30	19	9	2	6.19 - 6.17
bottom 30	8	15	7	0.29 - 0.00
total	51	29	10	1093.01 - 0.0

av. degree = 5.8

Table 8: Judgment on edges extracted from D

Note that the accuracy of the top 30 edges are very high. This suggests that the weight defined by (1) is appropriate. Next, over a half of the extracted edges are correct, though the connected components involve a lot of ambiguities, with the av. degree between 3.74 and 8.47. The authors are planning to improve the accuracy by revising the weight calculation and taking into account other information in the source ontologies. It should also be studied what happens if the notion of a match is generalized to include one-to-many and many-to-one correspondences. For the latter purpose, various graph algorithms will be useful as well as maximum weight matching algorithm is useful for extracting a one-to-one correspondence from G.

5 Related Works

EuroWordNet[6] and Penman, Pangloss Projects [Knight and Luk, 1994][7] also aim at building multilingual ontologies. EuroWordNet addresses development of a multilingual ontology for four European languages. The ontology will be made from existing ontologies. The method proposed above may be useful for connecting those ontologies. A work in Penman, Pangloss Projects auto-

[6] http://www.let.uva.nl/~ewn/

[7] http://www.isi.edu/natural-language/nlp-at-isi.html

matically merge WordNet and Longman's Dictionary of Contemporary English. Their merging problem can also be regarded as a maximum weight matching problem (though they did not do so), which demonstrates the generality of the formulation proposed in this paper.

[Ogino *et al.*, 1997] attempt to manually align upper parts of EDR and WordNet. The bottom-up alignment reported above will be improved by incorporating their result.

6 Conclusion

The paper first discussed the role of a large scale multilingual ontology in Global Document Annotation. Next it compared the distances among words in EDR Concept Classification Dictionary and in WordNet. The comparison suggests that the two ontologies are not very similar in encoded distances, and that alignment should improve the overall quality of the ontology. Finally, the paper reported on an attempt to align the two ontologies by maximum weight matching algorithm. The results of the alignment shows that over a half of the correspondences were correct even in highly ambiguous cases. Since the current method uses only small parts of the entire source ontologies, the alignment quality could be further improved by exploiting other information in the ontologies.

References

[Knight and Luk, 1994] K. Knight and S. K. Luk. Building a large-scale knowledge base for machine translation. In *Proceedings of AAAI '94*, 1994.

[Miller, 1995] George A. Miller. WordNet: A lexical database for English. *Communications of the ACM*, 38(11):39–41, 1995.

[Ogino *et al.*, 1997] Takano Ogino, Hideo Miyoshi, Masahiro Kobayasi Fumihito Nishino, and Jun'ichi Tsujii. An experiment on matching EDR Concept Classification Dictionary with WordNet. 1997.

[Yokoi, 1996] Toshio Yokoi. The EDR electronic dictionary. *Communications of the ACM*, 38(11):42–44, 1996.