ABSTRACT
We show the usefulness of textbooks in automatically forming a biomedical ontology by merging relationships between concepts (such as IS-A relations) picked out by clue expressions (e.g. ‘such as’, ‘called’). There are studies that attempt to make such an ontology from papers. However, it has been difficult to construct one complete structure.

Papers mainly contain relationships between a novel concept at the leaf level and a concept at an arbitrary middle level, and few relationships between such basic concepts which work as connectors for merging. By contrast, textbooks describe knowledge orderly, allowing us to construct with reliability a basic ontology which works as such connectors. If we use this ontology as a base structure when we merge the relationships extracted from papers, we can obtain a larger and more complete ontology. In the experiments, we compare the ontology constructed from both papers and textbooks with the ontology made only from papers.

1. INTRODUCTION
Domain ontologies are useful in various natural language processing tasks. For example, we can use ontologies as clues to query expansion. In other case, we can use ontologies in machine learning process to reduce data sparseness, and so on.

For the biomedical domain, some ontologies have been constructed by hand. However, such construction by human is very costly. Consequently, they often become out of date. Therefore, it is a desired alternative to construct ontologies from texts automatically and quickly. Though we need high quality ontologies to perform high quality information extraction or information retrieval, useless relationships are perhaps included in such automatically made ontologies and the quality of the ontologies becomes lower. In this paper, we use textbooks to improve the quality of the resulting ontology.

When we make ontology automatically, it is most likely to use text database, such as MEDLINE (Medical Literature, Analysis, and Retrieval System Online). From these texts, we acquire ontology by extracting relationships using the stable clue expressions (Hearst[7]) or the syntactic patterns (Pustejovsky et al.[11]) and by merging the relationships into structures based on their hierarchy (Rydin[13]). However, if we apply these methods directly to the ontology construction from biomedical papers, we might face two problems.

Firstly, the resulting ontology is too small to help us in various applications. Because the papers are written to represent discoveries and the author describes these discoveries by combining advanced concepts, we acquire just the relationships between these discoveries and the advanced concepts. If we acquire only these relationships, we cannot construct large structures, since we have no connectors that unite the advanced concepts. As a result, we obtain many small useless structures.

Secondly, the resulting ontology includes gaps. The author of the paper describes the discoveries connected to the basic concepts, like ‘protein’, ‘cell’, etc. Such relationships between the discoveries and the basic concepts have the large illogical jumps from the basic concepts to the discoveries. That is, because the authors do not write papers to explain step by step how to construct advanced concepts from basic concepts, we cannot acquire the missing links for those illogical jumps, even if we would construct the relationships completely. As a result, we become to have the ontology with large gaps. It is difficult to understand such an ontology, because the correct logic for filling the gap is unknown.

To deal with these problems, we focus on the properties of textbooks. In contrast to papers, textbooks describe knowledge from basic to advanced orderly. If we use textbooks, we can extract relationships step by step from basic ones to advanced ones. By using these relationships, we can con-
struct the ontology without gaps between concepts. We call such an ontology "the basic ontology". We can use this ontology as the foundation in constructing relationships extracted from papers. Because the basic ontology includes the continual relationships from basic to advance, the discoveries extracted from papers can be inserted into the proper position of the resulting ontology. Consequently, the resulting ontology of this construction will be larger and more ordered than that from only papers.

In this paper, we verify the above idea. For this purpose, we use biomedical paper abstracts (in MEDLINE database) and textbooks (provided by National Center for Biotechnology Information (NCBI)[1]). In order to extract relationships from these texts, we use clue expressions (e.g. ‘such as’, ‘called’). Then, we merge concepts indicated by the relationships. Since the relationships extracted from these clue expressions are mainly IS-A relationship, we construct these relationships into hierarchal structure, which we assume as the ontology.

In the experiments, we make two kinds of ontologies. One is made only from paper abstracts or only from textbooks. The other is made from both textbooks and paper abstracts. In the latter case, we first make basic ontology from textbooks and then insert relationships from paper abstracts into the ontology. We evaluate the quality of ontologies in two ways. First we compare the ontology only from abstracts with that from both textbooks and abstracts in terms of the node depth. Second we compare our ontologies with an existing ontology, Gene Ontology (GO)[5]. Experimental results show that the ontology constructed from both abstracts and textbooks is deeper and more hierarchal. Therefore we conclude that textbooks are useful to improve the quality of the resulting ontology.

In Section 2, we mention the related works. In Section 3, we precisely describe how to extract relationships from texts and how to construct those relationships into ontology. In Section 4, we describe the experiments and show the result. In Section 5, we draw the conclusion of this paper.

2. RELATED WORK

There are vast studies for extracting ontologies, not limited to biomedical domain (Blaschke et al.[3], Morin and Jacquemin[9], Rydin[13]). Moreover, there are also many studies for extracting various relationships (Rüdiger et al.[12], Khoo et al.[8], Craven and Kumlien[6]). However we have not yet seen the study in which the properties of using texts are considered or in which the order of using texts are considered.

Pustejovsky et al.[11] is similar to our study. They augmented UMLS[2] with supertype-subtype relationships extracted from MEDLINE. That is, they thought UMLS as the base ontology. However, in the biomedical domain, such basic knowledge still continues to change and expand. Therefore, for keeping the systems up to date, it is also important to construct the basic ontology automatically.

3. METHODS FOR MAKING ONTOLOGY

We make ontology from text in two steps. Firstly, we extract relationships between concepts. Secondly, we merge those relationships.

3.1 Extracting Relationships

In texts, there are some expressions which indicate relationships between concepts. We use such expressions as clues for extracting relationships and make a relational tree fragment.

In this paper, we use 'such as', 'e.g.', 'called', and 'like' as the clue expressions. Example sentences that include these clues are as follows:

- A mammal that starts as one cell becomes an organism with hundreds of diverse cell types such as muscle, nerve, and skin.
- One example is the synthesis of small peptides (e.g. glycylalanine) or proteins from amino acids.
- Each strand of DNA is composed of just four types of monomers called nucleotides.
- Single-celled organisms, like bacteria, have special outer coats to protect them.

The relationships indicated by these clues are mainly IS-A relations. We extract such relationships between concepts by using raw text, part-of-speech (POS) tagged text and parsed text.

In order to extract relationships, we first search clue expressions. When we find out sentences including those clues, we make sure that the preceding word of the clue is a noun. Additionally, in the case of 'like', we adopt the sentence only when comma appears just before the clue and the preceding word of that comma is a noun.

Next, we decide the part of the sentence where the clue indicates the relationship between concepts. The part is marked out by keywords. The keywords have two types: one for determining the start of the part and one representing the end of the part. Those keywords are determined by their POS tags. In Table 1, we show the keyword POSs.

Then we pick out noun phrases in that part because those noun phrases represent the concepts. By following bracket structures in the parsed text, we can find base phrase chunks in the text. In noun phrases, we distinguish head nouns and their modifiers. The head noun is the last word of the noun phrase. The modifiers of a head noun are the words tagged as 'nomin', 'adjective', 'adverb', 'participle', 'cardinal number', or 'predeterminer' in the same chunk or the preceding chunks without a head noun. Such non-head chunk is joined to the following noun phrase.

Lastly, we construct a basic relational tree from the noun phrases in the part. The noun phrase preceding the clue becomes the root node and the noun phrases following the clue become the leaf nodes.

In the part indicated by clue expressions, there are sometimes more than one noun phrases which preceding the clue. If the clue is 'e.g.', comma plus 'such as', or comma plus 'called', we simply adopt the noun phrase nearest to the
In the experiments, we used the following twelve textbooks: “C. elegance II”, “Cancer Medicine”, “Genes and Disease”, “Introduction to Genetic Analysis”, “Medical Microbiology”, “Modern Genetic Analysis”, “Molecular Biology of the Cell”, “Molecular Cell Biology”, “Retroviruses”, “The Human ATP-Binding Cassette (ABC) Transporter Superfamily”, “Surgical Treatment - Evidence Based and Problem-Oriented”, and the chapter 6 of “Vaccinia”.

We obtained biomedical abstracts from MEDLINE. MEDLINE is the U.S. National Library of Medicine’s (NLM) premier bibliographic database that contains over 12 million references to journal articles in life sciences with a concentration on biomedicine. NLM provides some ways to access MEDLINE and PubMed is one of those ways. Then we obtained 100 thousand references numbered as 19999999 on PubMed. The references contain title and abstract of the article. We picked out titles and abstracts from these references and used them in the experiments. However, some references do not contain title or abstract. In this case, we abandoned to use such references.

As described in the previous chapter, we used not only raw texts but also POS tagged texts and parsed texts. We used a POS tagger adapted for biomedical texts by using the GENIA corpus[10] as the training corpus. We also used Charniak’s statistical parser (ulpars[4]) to parse the texts.

4.2 Result

When we make an ontology from texts as described above, we obtain not only one large relational structure but also many small fragmented relational trees. Because such small trees are noisy, we adopt only the largest relational structure as ontology used for evaluation in the following sections.

4.2.1 Depth Distribution

In this experiment, we compare the appropriateness of the ontology constructed from both textbooks and abstracts with that only from abstracts, in terms of the node depth distribution of these ontologies. As we mentioned in Section 1, we assume that the resulting ontology from both textbooks and abstracts becomes deeper and more ordered than that from abstracts only.

It is hard to define the depth of the ontology nodes because the resulting ontology becomes a directed acyclic graph. Here we simply use the depth of the longest path. That is, although there are many routes from “root” nodes (nodes that have no parent node) to one node, we define the longest one of these routes as the longest path for the node. Length of a path is defined by the number of nodes included in the path.

We made the following ontologies:
Table 2: The characteristics of the ontologies

<table>
<thead>
<tr>
<th></th>
<th>nodes</th>
<th>maximum depth</th>
<th>average depth</th>
</tr>
</thead>
<tbody>
<tr>
<td>ontology-A</td>
<td>3122</td>
<td>19</td>
<td>3.85</td>
</tr>
<tr>
<td>ontology-T</td>
<td>2731</td>
<td>19</td>
<td>3.14</td>
</tr>
<tr>
<td>ontology-TA</td>
<td>4669</td>
<td>22</td>
<td>5.41</td>
</tr>
<tr>
<td>ontology-F</td>
<td>5314</td>
<td>22</td>
<td>7.66</td>
</tr>
</tbody>
</table>

‘nodes’ indicates the number of nodes included in each ontology. ‘maximum depth’ indicates the depth of the deepest node in each ontology. ‘average depth’ indicates the mean depth of all the nodes in each ontology.

ontology-A: an ontology made from 100 thousand abstracts
ontology-T: an ontology made from all the textbooks
ontology-TA: an ontology made from all the textbooks and part of abstracts (The amount of these texts is almost the same as ontology-A.)
ontology-F: an ontology made from all the texts, i.e., both all the textbooks and 100 thousand abstracts

We show the characteristics of these ontologies in Table 2. In addition, the amount of used text in ontology-A is about two times as large as that in ontology-T.

Considering from Table 2, as we supposed in the introduction, ontology-T (made only from textbooks) seems to have more basic relationships than ontology-A (made only from abstracts). Moreover, when we make ontology, the more abstracts we use together with textbooks, the deeper and larger the resulting ontology becomes.

Next, we show the detailed comparison, using ontology-A and ontology-TA because the amount of used texts in construction is nearly the same. Figure 1 shows the distribution of node depth for each ontology. As we see in Figure 1, the ratio of deep nodes in ontology-TA is greater than that in ontology-A. Therefore, we could insist that ontology-TA is larger and more ordered than ontology-A although the amount of source texts are almost the same.

4.2.2 Comparison with GO

Next, we evaluate the resulting ontologies by comparing them with an existing ontology. Here we use GO[5] as the existing ontology. In GO, 10979 terms are defined.

At first, we compared the depth distribution of GO and ontology-F. The results are shown in Figure 2. Although GO has the most nodes in depth 7, ontology-F has the most nodes in depth 3, a very shallow area. Ontology-F exceeds GO in the ratio of the nodes in the deep area.

Then we compare our ontologies with GO more precisely.

For this comparison, we examined ontologies in a node-by-node manner. However, if we define matching nodes as the nodes that contain the same phrase, we obtain a very few matching nodes. Hence we use partial matching. If a node of our ontology includes/is included in the node in GO,
Table 3: The extracted relationships and the number of matched nodes

<table>
<thead>
<tr>
<th>ontology</th>
<th>relationships</th>
<th>GO ≥ ours</th>
<th>ours ≥ GO</th>
</tr>
</thead>
<tbody>
<tr>
<td>ontology-A</td>
<td>4162</td>
<td>197</td>
<td>103</td>
</tr>
<tr>
<td>ontology-T</td>
<td>3653</td>
<td>267</td>
<td>136</td>
</tr>
<tr>
<td>ontology-TA</td>
<td>5717</td>
<td>262</td>
<td>158</td>
</tr>
<tr>
<td>ontology-F</td>
<td>7815</td>
<td>302</td>
<td>196</td>
</tr>
</tbody>
</table>

‘relationships’ indicates the number of extracted relationships in the construction of each ontology. ‘GO ≥ ours’ represents that the node in GO ‘includes’ the node in our ontology. ‘ours ≥ GO’ represents that the node in our ontology ‘includes’ the node in GO. These columns indicate the number of node pairs that satisfy the condition.

![Graph showing depth distribution of GO nodes matched with nodes of our ontology](image)

Table 4: Depth distribution table of the GO nodes matched with nodes of our ontology (in case GO nodes ‘includes’ our ontology node)

<table>
<thead>
<tr>
<th>depth</th>
<th>ontology-A</th>
<th>ontology-T</th>
<th>ontology-TA</th>
<th>ontology-F</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>4.9</td>
<td>2.4</td>
<td>3.1</td>
<td>1.9</td>
</tr>
<tr>
<td>4</td>
<td>32.0</td>
<td>31.7</td>
<td>30.6</td>
<td>31.6</td>
</tr>
<tr>
<td>5</td>
<td>15.5</td>
<td>12.7</td>
<td>13.8</td>
<td>12.7</td>
</tr>
<tr>
<td>6</td>
<td>15.5</td>
<td>20.6</td>
<td>19.9</td>
<td>18.4</td>
</tr>
<tr>
<td>7</td>
<td>17.5</td>
<td>17.5</td>
<td>17.3</td>
<td>18.4</td>
</tr>
<tr>
<td>8</td>
<td>17.5</td>
<td>17.5</td>
<td>17.3</td>
<td>18.4</td>
</tr>
<tr>
<td>9</td>
<td>3.9</td>
<td>3.2</td>
<td>4.1</td>
<td>5.1</td>
</tr>
<tr>
<td>10</td>
<td>2.9</td>
<td>1.6</td>
<td>2.0</td>
<td>1.3</td>
</tr>
<tr>
<td>11</td>
<td>0</td>
<td>0.8</td>
<td>0.5</td>
<td>0.6</td>
</tr>
<tr>
<td>12</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

This table represents the same depth distribution as Figure 3. For each ontology, we show the percentage of nodes in each depth.

In ontology-A. Considering that the amount of texts used in constructing ontology-A is about two times as large as that in ontology-T, we can insist that we can extract concept relationships more effectively from textbooks than from abstracts. Therefore, if we use textbooks, we can construct ontologies faster from smaller texts than the case we use abstracts.

We now show the details of the comparison by the depth distribution of GO nodes. We determined the depth of GO nodes of the matched nodes as described in section 4.2.1 and made distribution maps (Figure 3 and Figure 4, cf. Table 4 and Table 5).

By comparing the graph of ontology-A with that of ontology-T, we can insist that the distribution of the concepts extracted from textbooks and that from abstracts are different. Especially, textbooks contain more nodes in the moderate depths.

5. CONCLUSION

In this paper, we proposed the idea that textbooks are useful to improve the quality of the automatically constructed ontology because they contain many basic relations which are not contained in papers and should work as connectors for abstract discoveries.

When we matched nodes in Section 4.2.2, we could not obtain sufficient number of the exact match pairs for evaluation. We then substituted manually partial match pairs. Even though we relaxed the criteria, we could not obtain sufficient pairs for evaluation based on their relationships in ontologies. Because such relationships and matched pairs are based on proper ontology (GO), we can examine the property of our ontologies if we have those relationships and pairs sufficiently. Therefore, for the present, we cannot discuss about our assumption that we described in the introduction because of the lack of evidence for this discussion.
These matters are remained as future work.

In spite of such problems, at least we can insist that textbooks are in some degree more effective than abstracts. Furthermore, as we mentioned in Section 4.2.1, the depth distribution of resulting ontology changes toward higher quality by using textbooks, as we expected.

Moreover, from the depth distribution comparison of GO and our ontology, we also see that the depth distributions of our ontology and GO are different. The peak difference of distribution might represent that more texts are required for making proper ontology. But, at the same time, our ontology exceeds GO in the ratio of the nodes in deep area. This indicates that our ontology could be better than GO in some points.

6. REFERENCES


