Toward Medical Ontology using Natural Language Processing

Eiji Aramaki†, Takeshi Imai†, Masayo Kashiwagi†,
Masayuki Kajino‡, Kengo Miyo†and Kazuhiko Ohe†

†Dept. of Planning, Information and Management, University of Tokyo Hospital
‡Japan Research Group for Medical Ontology
{aramaki,ken,masayo,kohe}@hcc.h.u-tokyo.ac.jp
kajino@medical-ontology.jp
miyo-tky@h.u-tokyo.ac

Abstract

In this paper, we introduce our project aiming to build a medical ontology, and also present a method to estimate term relations and term classification, which are the basic structure for the ontology. First, relations between medical terms are extracted from a medical electronic dictionary. Next, the terms are classified based on the co-occurrence verbs. Preliminary experimental results show the basic feasibility of our approach.

1 Introduction

So far, ontology is defined in various ways. In this paper, we define ontology as the following definition (Musen, 1998):

> An ontology is a model of a particular field of knowledge – the concepts and their attributes, as well as the relationships between the concepts.

Especially, in the medical field, a special attention is given to ontology studies(Stead et al., 2000; Musen, 2002; Hajdukiewicz et al., 2001), because a medical ontology can be a core technology for various applications, such as an electronic medical record, a diagnostic support system and so on(Musen et al., 1999; Sowa, 2000).

In addition, medical terms are strongly controlled, leading to strict usage and less ambiguity. This can be a strong advantage for a precise natural language processing.

Ave. # of sentences	4.37
Ave. # of char. per sentence	42.6

Table 1. Statistics of Normal Entries.

Entry Type	# of Entries
Normal Entry	48,229 (69.3%)
Synonymous Entry	6,443 (9.3%)
Reference Entry	14,932 (21.5%)

Table 2. The Number of each Entry Type.

However, most of ontology studies in the medical field mainly rely on hand-made approaches (SNOMED, 2002; UMLS, 2003). Such an approach suffers from maintenance costs.

In such a situation, our project has started in October 2004. The goal of the project is to build a Japanese medical ontology semi-automatically.

For the first stage of the project, we estimate the medical term relations and classification, which can be the basic structure of our ontology.

First, the system parses the lead (top) sentence of each entry in a medical dictionary. A lead sentence usually explains compactly the meaning of the entry, using its hypernym. Therefore, the system can extract the relations between an entry and its hypernym with a high-accuracy.

After that, the system classifies terms based on the distributional hypothesis, which states that words that occurred in the same contexts tend to have similar meanings (Harris, 1985).

The point of proposed method is the combination of the technique of the hyponyms extraction

(which has high-accuracy) and the term classification (which has high-coverage). Preliminary experimental results show the basic feasibility of the proposed approach.

This paper is organized as follows. The next section introduces our corpus. Section 3 describes the proposed method. Then, Section 4 reports experimental results, Section 5 describes related works, and Section 6 presents our conclusions.

2 Corpus/Materials

First of all, this section describes our corpus/material. We used Igakusyoin's medical dictionary(Ito et al., 2003), which is a standard medical dictionary in Japan. It consists of 69,604 entry terms, which are classified into the following three types:

1. **Normal Entry**: it explains the entry terms as follows.

   Early Stomach Cancer:
   The Early Stomach Cancer is a stomach cancer in which the carcinoma infiltration extend through the bottom of a gastric mucosa. It also may extend through the stomach wall and spread to nearby lymph nodes and to organs such as the liver, pancreas ...

   The statistics are shown in Table 1.

2. **Synonymous Entry**: This entry type only refers to its synonymous term. An example is as follows:

   Catabolic Action:
   =catabolism

3. **Reference Entry**: This entry type refers to its reference term as follows:

   Computer Game Epilepsy:
   →television epilepsy

The number of each type is shown in Table 2.

Both synonymous entries and reference entries can be directly converted into term relations. So, the proposed method focuses on handling normal entries.

3 Method

This section explains the proposed method, consisting of two steps.

Pattern	Extracted Relation
X is found as Y .	X–(hypernym)→Y
X is one of Y .	X–(hypernym)→Y
X is a part of Y .	X–(hypernym)→Y
basically X means Y .	X–(synonymous)→Y
X is named Y .	X–(synonymous)→Y
X is called Y .	X–(synonymous)→Y

Table 3. Patterns for Hypernym/Synonym Extraction.

* In fact, the patterns are described in the tree forms, but for simplicity, this table shows them without their structures.

First, the proposed method extracts term relations, mainly hypernym relations (STEP1). Then, the method classifies terms by using the STEP1 results (STEP2).

3.1 STEP1: Term Relation Extraction

The lead sentence in a dictionary description compactly explains the meaning of the entry. Especially, in the Japanese dictionary style, a head-word of a lead sentence usually indicates a hypernym of its entry. Therefore, we can easily extract the relations between an entry and its hypernym (Tsurumaru et al., 1991).

For example, a term "*Early Stomach Cancer*" is explained with its head-word "*Stomach Cancer*" as follows[1].

> *An Early Stomach Cancer* **is a stomach cancer** in which...

However, there are some exceptions. For example, the following entry "a residual stomach" has the head-word "*a part (of)*", and it is meaningless as the hypernym.

> *A Residual Stomach* **is a part** of stomach which ...

To deal with such exceptions, we prepared 46 patterns/rules for hypernym (or synonym[2]) extraction. Some examples of them are shown in Table 3.

The extraction algorithm is as follows.

First, for each normal entry, the system ana-

[1]All examples in this paper are translations in English.

[2]As shown in the table, some patterns do not extract hypernyms but synonyms.

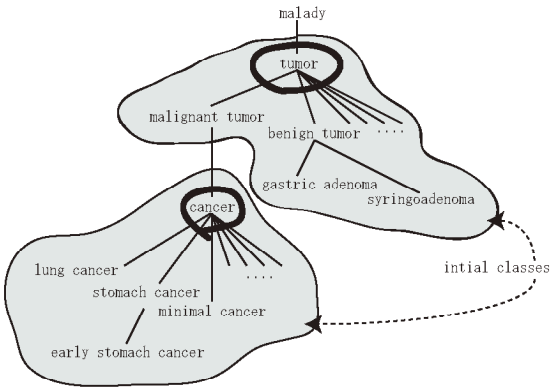


Figure 1. Examples of Term Relations.
* "An initial class" is explained in Section 3.2.

lyzes the lead sentence by Japanese parser KNP (Kurohashi and Nagao, 1994). If the parsing result matches a pattern, the system extracts the hypernym/synonym candidate by using the pattern. Otherwise, the system regards its head-word as a hypernym candidate.

Then, the system checks whether the candidate is suitable or not, by consulting the medical dictionary. In the case that the extracted hypernym/synonymous does not appear in the entries of the dictionary, the system rejects the relation (such entries have no relation).

After this procedure, the system sums up all relations, and obtains tree structures (as shown in Figure 1).

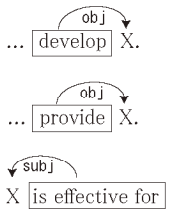
3.2 STEP2: Term Classification

The above procedure estimates reliable relations between terms. However, as mentioned before, if a relation is rejected, the term remains without relations. In order to handle such terms, the system classifies them in this step.

We conduct term classification based on the distributional hypothesis, which states that words that occurred in the same contexts tend to have similar meanings (Harris, 1985).

In applying the hypothesis, we limit a context of a term to its co-occurrence verb(v) and the syntactic relation(rel) (such as subject relations, object relations and so on[3]).

Such small context is workable because a term often appears with its typical verb as follows:



From the above three phrases, we can easily make a guess that X indicates a kind of a medicine (Of course, we limit X to a medical term).

Probabilistic Model

For the statistical formalization of the distributional hypothesis, we assume the following probabilistic model.

$$P(v, rel, n) = \sum_{c \in C} P(c)P(n|c)P(v, rel|c),$$

where n is a term, v is n's co-occurrence verb, rel is a relation between n and v, c is a class of n, C is a set of c.

Roughly speaking, this model means that a term appearance is tied into a verb and its relation appearance by a hidden class c.

Parameter Estimation with EM Algorithm

Under the above probabilistic model, all we have to do is to estimate three probabilistic tables, $P(v, rel|c)$, $P(n|c)$ and $P(c)$. The problem of the estimation is unobserved c (If c is observed, we can easily get these probabilities by direct counting.). To estimate three probabilities, we utilize an Expectation Maximization(EM) algorithm, which estimates unobserved parameters by an iterative procedure. See (Rooth et al., 1999) or (Torisawa, 2002) for more detail.

Initial Parameter Setting

EM-based word classification demonstrated its effectiveness in previous studies (Rooth et al., 1999; Torisawa, 2002). However, they used huge corpus, i.e., web data, newspapers and so on.

[3]A Japanese syntactic relation is analyzed by the Japanese parser KNP. We used the case-marker as is.

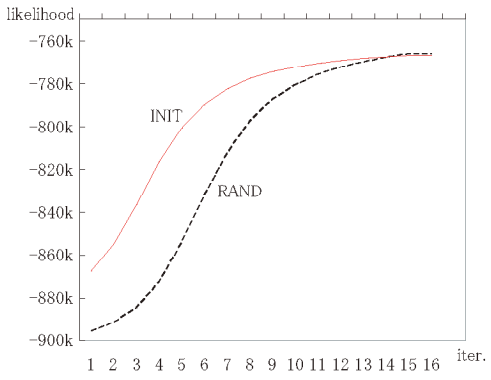


Figure 2. Likelihood and Iteration.

On the other hand, our corpus size is small as compared to them. The lack of data sometimes bring a local minimum problem to the algorithm. To avoid the problem, we give the algorithm rough categorization, which is represented as a set of initial probabilities by the following method.

First, we regarded a term which is the hypernym of many (more than a threshold[4]) terms as a local center. We also think that the center and its descendants make a category. We call such a category as an **initial class**[5].

For example, in Figure 1, we regard "*tumor*" and "*cancer*", which have many children, as initial classes.

After the initial class construction, initial probabilities are set by using the initial classes as follows. If a term n_i belongs to an initial class C_j, the initial probability $P(n_i|C_j)$ is $\frac{1}{|C_j|}$, otherwise $P(n_i|C_j) = 0$. Where $|C_j|$ denotes the number of C_j members.

4 Experiments

We think that the experiments should be done in the view point of an application performance. However, we have not developed a practical application yet. So, this paper only reports the statistics of each step.

[4]In the experiments in Section 5, we set the threshold to 40.

[5]In the case that an entry has two or more centers in its ancestors, we think that it belongs to the most recent ancestor's class.

Term	Extracted Relation & Term
E2F transcription factor	→ transcription factor
E5-1 gene	= presenilin 2
subtotal gastrectomy	→ operation
ERCC gene	→ gene
ERCP-induced pancreatitis	→ induced pancreatitis
gastric juice test	→ test
EMG syndrome	→ syndrome
chlorosis	→ anemia
ioxaglic acid	→ contrast medium
sodium iopodate	→ contrast medium

Table 4. Examples of Extracted Terms and Relations.

* Where "→" is a hypernym relation, "=" is a synonymous relation.

4.1 Term Relations

Experimental Setting

After STEP1, 25815 hyp/syn relations are extracted from the dictionary. They are more than half of the normal entries.

Evaluation

In order to check the validity of these relations, 100 relations are randomly extracted, and judged by four humans (two medical doctors and two nurses). They judged the hyp/syn relations whether valid or not.

Results

The average accuracy is high, 89.3% (min=84.0%; max=95.0%; average concordance rate is 91.7%). We think that it is because medical terms have less ambiguity, leading to the high performance. Some examples are shown in Table 4.

4.2 Term Classifications

Experimental Setting

We extracted 60,000 phrases (v, rel, n) from the dictionary, and classified them by the EM-based method in STEP2.

The method requires two parameters, the number of classes ($|C|$) and the number of iteration($iter$). We set $|C|$ to 100, and $iter$ to 16 after preliminary experiments.

Evaluation

After the classification, we extracted 50 terms (5 classes × 10 terms). Then, they are judged by four humans (three medical doctors and one nurse)

Method		Acuracy
RAND		51.0% (= 102/200)
INIT	(a+b)	84.0% (= 168/200)
	(a)	97.0% (= 97/100)
	(b)	70.0% (= 70/100)

Table 5. Term Classification Accuracy.

* INIT terms are classified into two types; (a) the terms with the initial classes and (b) newly classified terms. We balanced the number of (a) and (b).

whether a term can be suitable as a member of its class [6].

To compare the validity of the initial classes (mentioned in Senction 3.2), we conduct two methods as follows:

RAND: initial probabilities are set to random values.

INIT: initial probabilities are set by using initial classes as mentioned in Section 3.2.

Results

The result is shown in Table 5. As shown in the table, INIT demonstrated higher accuracy than RAND. However, we have to say that it is an ambiguous task for human, because the average concordance rate between humans is only 71%.

Figure 2 shows the likelihoods of each iteration step. Not surprisingly, INIT jumps up its likelihood in early steps. However, in the last steps, both INIT and RAND have almost equal likelihood. It might indicate that the likelihood does not reflect the classification validness.

Table 6 shows some examples of the INIT classification.

4.3 Discussion

As I've mentioned before, we need an application to investigate the validity of our ontology. However, we believe that the proposed method is promising because of the high accuracy in the preliminary experiments.

[6]Because the class label/name is implicit, human evaluaters guessed the class name, and then judged the validness.

n	$P(n\|c_{17})$
joint	0.131
inferior limb	0.089
extremity	0.061
digit	0.043
head	0.034
hip joint	0.031
upper extremity	0.025
ankle joint	0.019
foot	0.018
forearm	0.018
wrist joint	0.016
knee	0.015
vertebral column	0.012
:	:
v (rel)	$P(v, rel\|c_{17})$
flex (obj)	0.124
set (obj)	0.057
is a degeneration (of)	0.051
move (obj)	0.049
consist (of)	0.040
crook (obj)	0.035
locate (in)	0.034
is near (of)	0.021
is a part (of) ...	0.015
is an anomaly (of)	0.014
(subj) is stable	0.011
:	:

Table 6. Examples of Class 17 Probabilities.

* **A bold term** is a member of an initial class "joint".

5 Related Works

In the field of medical informatics, many studies are conducted for ontology, and some of them realized practical results, i.e., (SNOMED, 2002; UMLS, 2003) and so on. However, as mentioned before, they rely on mainly hand-made approaches, compeling huge maintenance costs.

On the other hand, in the field of natural language processing, many studies pay attention to full/semi-automatical techniques for knowledge-base building. For example, (Richardson et al., 1998) built MINDNET, which is a syntactic knowledge-base using an encyclopedia. (Tsurumaru et al., 1991) build a thesaurus using a Japanese word dictionary. Our approach is similar to their approach. The main deference is that we incorporate term classification into their techniques.

6 Conclusions

In this paper, we introduced our medical ontology project, and proposed a method to estimate medical term relations/classifications. Our techniques are

based on the hypernym extraction and the distributional hypothesis. Preliminary experimental results showed basic feasibility of our approach.

Acknowledgments

Part of this research is supported by Grant-in-Aid for Scientific Research(A) of Japan Society for the Promotion of Science iProject Number:16200039, F.Y.2004-2007j and the Research Collaboration Projcct with Japan Anatomy Laboratory Co.Ltd.

References

J. R. Hajdukiewicz, K. J. Vicente, D. J. Doyle, P. Milgram, and C. M. Burns. 2001. Modeling a medical environment: an ontology for integrated medical informatics design. *Int.J.Medical Informatics*, 62(1):79–99.

Z. Harris. 1985. Distributional structure. In *The Philosophy of Linguistics, Oxford University Press*, pages 26–47.

M. Ito, H. Imura, and H. Takahisa. 2003. *IGAKUSHOIN'S MEDICAL DICTIONARY*. Igakusyoin.

Sadao Kurohashi and Makoto Nagao. 1994. A syntactic analysis method of long Japanese sentences based on the detection of conjunctive structures. *Computational Linguistics*, 20(4).

M. A. Musen, S. W. Tu, A. K. Das, and Y. Shahar. 1999. EON: A component-based approach to automation of protocol-directed therapy. *Journal of the American Medical Informatics Association*, 3(6):367–388.

M. A. Musen. 1998. Domain ontologies in software engineering: use of protege with the EON architecture. *Methods of Information in Medicine*, 37(1):540–550.

M. A. Musen. 2002. Medical informatics: Searching for underlying components. *Methods Inf. Med.*, 41(1):12–19.

S. D. Richardson, W. B. Dolan, and L. Vanderwende. 1998. MindNet: acquiring and structuring semantic information from text. In *Proceedings of 36th Annual Meeting of the Association for Computational Linguistics (ACL) and the 17th International Conference on Computational Linguistics (COLING)*, pages 1098–1102.

Mats Rooth, Stefan Riezler, Detlef Prescher, Glenn Carroll, and Franz Beil. 1999. Inducing a semantically annotated lexicon via EM-based clustering. In *Proceedings of the ACL 1999*, pages 104–111.

SNOMED. 2002. *SNOMED Clinical Terms Guide*. College of American Pathologists.

J. F. Sowa. 2000. *Knowledge Representation: Logical, Philosophical, and Computational Foundations*. Brooks/Cole Publishing Co. Pacific Grove, CA.

W. W. Stead, R. A. Miller, M. A. Musen, and W. R. Hersh. 2000. Integration and beyond: Linking information from disparate sources andinto workflow. *J. Am. Med. Inform. Assoc.*, 7(2):135–145.

Kentaro Torisawa. 2002. An unsupervised learning method for associative relationships between verb phrases. In *Proceedings of 19th International Conference on Computational Linguistics (COLING)*, pages 1009–1015.

Hiroaki Tsurumaru, Katsunori Takesita, Katsuki Itami, Toshihide Yanagawa, and Sho Yoshida. 1991. An approach to thesaurus construction from Japanese language dictionary (in Japansese). *Natural Language Proceeding*, (16).

UMLS. 2003. *UMLS KNOWLEDGE SOURCES, 14th Edition*. National Institutes of Helth Department of Helth and Human Services: U.S. National Library of Medicine.