SERVING KNOWLEDGE RESOURCES WITH ONTOLOGIES

Takehiro Tokuda[†], *Takenobu Tokunaga*[†] *and Akifumi Tokosumi*[‡]

[†]Department of Computer Science, [‡]Department of Value and Decision Science
Tokyo Institute of Technology

ABSTRACT

Ontologies are widely used in various research and engineering areas. We present our three experiments to deal with knowledge resources using ontologies.

1. INTRODUCTION

Ontologies are widely used in various research and engineering areas such as natural language processing, information retrieval, database design, artificial intelligence, the Semantic Web, Web services, software system design, and cognitive sciences.

Possible definitions of ontologies may be as follows.

0. A collection of terms
1. A collection of terms and their various relations (at least including the is-a relation)
2. A collection of terms, their properties, and their various relations
3. A collection of terms with definitions and various relations
4. A collection of terms, their properties, and machine-executable relations
5. A collection of terms with machine-executable definitions and relations

The 0th definition is the weakest one used for writing of documents by a group of people. The 1st and 3rd definitions are used for natural language ontologies. The 2nd definition is used in database design and software system design. The 4th definition is used in Web ontologies for the Semantic Web and Web services. The 5th definition is the strongest one which may be considered as the level of knowledge representations (higher than the level of ontologies).

Ontologies may be useful to share a framework of knowledge among humans and machines. Using a well-designed ontology, we could access instances easily according to known properties of those instances. We also could understand the relationships of two instances according to their categories.

Ontologies may, however, have some possible limitations. One universal ontology of everything seems very difficult to construct. Each application tends to create its own ontologies. We need to map one ontology to another. In the case of automatic handling of ontologies, expressions of relationships are strongly limited by computational decidability.

The purpose of this paper is to present our three experiments to deal with knowledge resources using ontologies. Our first experiment is to construct a large-scale computer science text corpus and improve accesses to necessary sample sentences using an ontology of intentions. Our second experiment is to map an English ontology to a Japanese ontology. Our third experiment is to construct an ontology of Knowledge/Affective Resources

The organization of the rest of this paper is as follows. In Section 2, 3, and 4, we respectively explain out first, second, and third experiments respectively. In Section 5 we give concluding remarks.

2. A LARGE-SCALE COMPUTER SCIENCE TEXT CORPUS AND QUERY BY INTENTIONS

We describe a large-scale computer science text corpus called X-Tec and its query method using intentions.

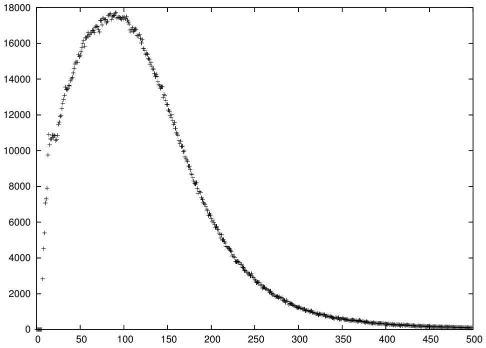
Yusuke Soyama and Takehiro Tokuda constructed a large-scale computer science text corpus. Our text corpus has 2.98 million sample sentences and their original URLs, which are collected from English computer science papers in PDF files on the Web.

Our X-Tec system consists of four subsystems: a crawler subsystem, X-Tec databases, a query subsystem, and an expression diagnostics subsystem. The crawler subsystem collects computer science text corpus information from the Web. The task of our crawler is same as that of the Web search engine crawlers except the recognition of the level of one sentence.

Fig. 1 shows the distribution of the byte length of English sentences in our X-Tec databases. The total number of sentences is 2,978,612. The total byte length of sentences is 357,183,201. The average byte length of one sentence is 119.92.

Our query subsystem allows us to search for sample sentences and URLs by using ordinary conditions such as oc-

Fig. 1. A distribution of the length of the sentences



currences of one or more words, distances of words, derived words, and parts of speech.

Our expression diagnostics subsystem allows us to automatically diagnose expressions of a given paper using frequencies of unigrams, bigrams, and skipped bigrams of verbs, nouns, and prepositions.

In addition to ordinary queries, we introduced query by intentions. We constructed an ontology of locations inside the paper and intentions of expressions so that we may be able to retrieve necessary sample sentences by intentions.

Our query by intentions can be processed as follows.

(1) A pair of the location inside the paper and the intention of the expressions is selected by a user.

(2) For the selected pair of the location and the intention, a number of words are associated. For example, “motivation”, “background”, “inspired” are words associated with the motivation explanation in the Introduction section.

(3) Sample sentences are collected using these associated words. The resulting sample sentences are shown according to various orderings.

At present the query subsystem provides valuable information about sample sentences which cannot be obtained by general English corpora [1, 5] or general Web search engines [2]. For example, our corpus shows many appropriate verbs for the noun “conclusion” in the computer science papers. Our corpus also shows exact frequencies of “stand-alone”, “standalone”, and “stand alone” respectively as 300, 195, and 43.

The expression diagnostics subsystem is not yet performing well, maybe because the number of sentences is not yet large enough. A query method by intentions is in the stage of preliminary experiments. We hope this query method will improve the accessibility.

3. MAPPING BETWEEN UPPER LEVEL ONTOLOGIES

The importance of ontologies has been widely recognized in various application domains, and various ontologies have been constructed in each domain and each language. However, it is also well known that constructing a universal wide-coverage ontology is still a big challenge. One of the approaches to tackle this problem is to construct a universal ontology covering only upper level and to specialize it to a specific domain and language as needed [6, 8]. This approach assumes that it is possible to build a consensus on upper level structure of ontologies. In the light of such background, this section poses the following questions and try to answer them through preliminary experiments.

- Is it really possible to build a consensus on upper level structure of ontologies?
- Does language affect on upper level structure?
- Is it possible to make mapping between ontologies automatically?

To answer these questions, we tried to make mapping between SUMO (The Suggested Upper Merged Ontology) [6] and a Japanese thesaurus *Nihongo Goi Taikei* [4]. SUMO is an upper level ontology that has been proposed as a starting point for The Standard Upper Ontology Working Group, an IEEE-sanctioned working group, defining about 700 concepts together with more than 2,000 assertions on the concepts. SUMO uses English for documentation of concepts. *Nihongo Goi Taikei* classifies a total of 264,312 nouns into 2,710 semantic classes with a hierarchical structure.

We first tried to make mapping between SUMO concepts and *Nihongo Goi Taikei*’s semantic classes manually by referring to descriptions of SUMO concepts and the structure of *Nihongo Goi Taikei*. We focused on mapping of noun concepts, that is, we tried to make mapping between 630 SUMO concepts and 2710 *Nihongo Goi Taikei* semantic classes.

For example, a SUMO concept “AttachingDevice” is mapped to a *Nihongo Goi Taikei*’s semantic class “*950 Sagyôgu (Setuzoku)* (connecting parts)”, since the documentation of “AttachingDevice” is “A &%Device whose purpose is to attach one thing to something else, e.g. nails, screws, buttons, etc.” and semantic class “*Sagyôgu (Setuzoku)*” includes words“*kugi* (nail)”, “*nezi* (screw)” and “*botan* (button)”[1]. As shown in this example, lexical information provides important clue for making mapping.

Among 630 SUMO concepts, we succeeded to assign synonymous *Nihongo Goi Taikei*’s semantic classes to 289

[1] A symbol “&%” denotes a SUMO concept, and a preceding number of a semantic class denotes its identifier.

SUMO concepts. There were cases in which a SUMO concept is mapped to more than one *Nihongo Goi Taikei*'s semantic classes and vice versa, the number of synonymous relations became 282. There were 217 SUMO concepts which had no synonymous semantic class but hyponymous (209) or hyperonymous classes (8). 124 SUMO concepts had no counterpart in *Nihongo Goi Taikei*, which is equivalent of 19.7% of all the SUMO concepts.

There were several types of problem in mapping between SUMO and *Nihongo Goi Taikei*. First, there are differences in viewpoints of classifying subconcepts/subclasses. For example, a SUMO concept "GeographicArea" has "LandArea" and "WaterArea" as its subconcepts. On the other hand, a *Nihongo Goi Taikei*'s semantic class "*tikei* (land form)" corresponding to "GeographicArea" are further classified into "*rikuti* (land)" and "*umi* (sea)". Thus, "River-*kawa* (river)" would be classified under "WaterArea" in SUMO but under "*rikuti* (land)" in *Nihongo Goi Taikei*. This might be called horizontal mismatch.

Second, a leaf concept has a corresponding semantic class which has further subclasses, and vice versa. For example, SUMO classfies "PhysicalQuantity" into "ConstantQuantity", "UnitOfMeasure" and "FunctionQuantity", but *Nihongo Goi Taikei* does not classify "*tanni* (measure)", counterpart of "PhysicalQuantity", into further subclasses. This might be called vertical mismatch or granularity problem.

Third, a concept and a corresponding semantic class have entirely different substructures. This could be an extreme case of the horizontal mismatch. For example, a SUMO concept "Relation" corresponds to a *Nihongo Goi Taikei*'s semantic class "*kanren* (relation)". Although there is quite a lot of correspondence between subconcepts and subclasses under them, the structure is quite different.

Some of cases of these problems come from a deep philosophical basis, and difficult to solve. Others, however, could be solved by simple operation. For example, merging a set of subconcepts and subclasses could solve the horizontal mismatch for some cases.

Next, we tried to make automatic mapping on the basis of the number of overlapping words of a SUMO concept and *Nihongo Goi Taikei*'s semantic class. Since SUMO itself has no lexical information, we used linking resources [7] between SUMO and WordNet [3], a large English lexical database, to incorporate lexical information into SUMO. Associated English words in a SUMO concept were translated into Japanese words by using EDICT[2], a free Japanese-English dictionry including more than 100,000 entries. The number of words associated with both a SUMO concept and a *Nihongo Goi Taikei*'s semantic class were used as mapping score. We assigned a *Nihongo Goi Taikei*'s semantic class with the highest score to each SUMO concept. As a results, we could make mapping for 279 SUMO concpts automatically, and 154 of them were concepts which could be also manually assigned *Nihongo Goi Taikei*'s semantic classes. Among these 154 concepts, 64 cases were consistent with the manual mapping. That is, the accuracy of automatic mapping is 23% (64/279). We further analyzed 125 cases which semantic classes could not be assigned manually but could be by machine. We found hyponymous or hyperonymous relations in 74 cases and synonymous relations in 5 cases. This suggests that automatic mapping would help manual mapping

This section started with questions on mapping upper level ontologies in different languages, and described preliminary experiments to answer these questions. In summary, there is quite a lot of differences between even upper level ontologies, but automatic mapping could be possible with human assistance.

4. LANGUAGE AND ONTOLOGY FOR CRITICS

When people speak about a certain thing, it will sometimes "attract" a particular class of words. Classical music, for instance, is often described in terms of the geographical location from which the music originates as well as musicological terms that categorize the piece. Electronica (electronic pop music usually created with a computerized sequencer), on the other hand, tends to attract terms that specify the musical style, genre, and the industry.

To obtain insights into the problem, a quantitative analysis of music reviews for all major genre of music was conducted [9], which collected 2,439 CD reviews from 11 Japanese music review magazines published in 2004. A morphological analysis identified 236,220 words, from which just 42,477 content words (nouns, adjectives, adverbs, and verbs) were used for analysis.

The reviews were classified into 19 music genres with nine upper-level genre categories (1:[world, reggae, folk, new-age], 2:[independent-rock, post-rock, avant-garde], 3:[hip-hop, RB], 4:[house, techno], 5:[classical, soundtrack], 6:[electronica, break-beats], 7:[mainstream-rock, pop], 8:[Japanese-pop], 9:[jazz]). The content words in each music genre were then classified into the 19 semantic categories, denoting music related concepts (music structure, instrument, genre, style, player, industry, etc.) and critical comments (evaluation, sensibility, geography, time, etc.). Dependencies between music genre and semantic category were analyzed.

From the significant dependencies observed, the following points may be made; (i) in genre 9:[jazz], player-related words are more frequent than expectancy levels, (ii) in genre 1:[world, raggae, folk, new-age], geography, player, genre-

[2] http://www.csse.monash.edu.au/~jwb/j_edict.html

related words are more frequent than expectancy levels, (iii) in genre 2:[independent-rock, post-rock, avant-garde], abstract descriptions on the music are dominant, while (iv) in genre 7:[mainstream-rock, pop], player, music structure, genre-related words are dominant.

Each music genre, with its own musical, social, and historical constraints, has its own unique form of discourse. These observations suggest that a music ontology alone is not necessarily sufficient to enable a real person/machine to speak about a piece of music. We propose two possible solutions and will discuss their merits:

- specialized ontologies for every local segment of the target field (e.g. the genre of music)
- combination of a general ontology (e.g. music in general) and specialized ontologies (e.g. the genre of music)

5. CONCLUSION

We have presented our three experiments to handle knowledge resources using ontologies. Ontologies seem one of important methods to deal with large-scale knowledge resources.

6. REFERENCES

[1] British national corpus. `http://www.natcorp.ox.ac.uk`.

[2] Google. `http://www.google.com`.

[3] C. Fellbaum. *WordNet: An Electronic Lexical Database*. The MIT press, 1998.

[4] S. Ikehara, M. Miyazaki, A. Yokoo, H. Shirai, S. Nakaiwa, K. Ogura, Y. Ooyama, and Y. Hayashi. *Nihongo Goi Taikei – A Japanese Lexicon*. Iwanami Syoten, 1997. (In Japanese).

[5] D. Lea, editor. *Oxford Collocations Dictionary for Students of English*. Oxford University Press, 2002.

[6] I. Niles and A. Pease. Towards a standard upper ontology. In *Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001)*, 2001.

[7] I. Niles and A. Pease. Linking lexicons and ontologies: Mapping WordNet to the subbested upper merged ontoogy. In *Proceedings of the 2003 International Conference on Information and Knowledge Engineering (IKE 03)*, 2003.

[8] P. Vossen, editor. *Euro WordNet*. Kluwer Academic Publishers, 1998.

[9] D. Moriyasu, and A. Tokosumi. *Constructing an Ontology of Music Criticism.* forthcoming.