Acquiring Conceptual Relationships from Domain-Specific Texts

Takahira Yamaguchi
Dept. Computer Science, Shizuoka University
3–5–1 Johoku Hamamatsu Shizuoka 432-8011 JAPAN
yamaguti@shizuoka.ac.jp

Abstract

Here is discussed how to construct domain ontologies with both taxonomic and non-taxonomic conceptual relationships, exploiting a machine-readable dictionary (MRD) and domain-specific texts. The taxonomic relationships come from WordNet in the interaction with a domain expert, using the following two strategies: match result analysis and trimmed result analysis. The non-taxonomic relationships come from domain-specific texts with the analysis of lexical co-occurrence statistics; based on WordSpace to represent lexical items according to how semantically close they are to one another. We have done case studies in the field of law. The empirical results show us that our environment can support a user in constructing domain ontologies.

1 Introduction

Although ontologies have been so popular in many application areas, we still face the problem that it takes many costs to build up them by hand. In particular, since domain ontologies have the senses specific to application domains, human experts have to make huge efforts with constructing them entirely by hand.

In order to reduce the costs, automatic or semi-automatic methods have been proposed using knowledge engineering techniques and natural language processing ones (cf. Ontosaurus[Swartout et. al. 1996]). The authors have also developed a domain ontology refinement support environment called LODE[Kurematsu and Yamaguchi 1997] and a domain ontology rapid development environment called DODDLE[Sekiuchi et. al. 1998], using machine readable dictionaries. However, these environments facilitate the construction of just a hierarchically structured set of domain concepts, in other words, taxonomic conceptual relationships.

As domain ontologies have been applied to widespread areas, such as knowledge sharing, knowledge reuse, software agents and information integration, we need software environments that support a human expert in constructing the domain ontologies with not only taxonomic conceptual relationships but also non-taxonomic ones. In order to develop the environments, it seems to be better that we put together two or more techniques such a knowledge engineering, natural language processing, machine learning and data engineering, as seen in the workshop on ontology learning in ECAI2000 (e.g. [Maedche and Staab 2000]).

Here in this paper, we extend DODDLE into DODDLE II that acquires both taxonomic and non-taxonomic conceptual relationships, exploiting WordNet[Fellbaum 1998] and domain-specific texts with the automatic analysis of lexical co-occurrence statistics, based on WordSpace that has the idea that a pair of terms with high frequency on co-occurrence statistics can have non-taxonomic conceptual relationships. Furthermore, we evaluate how DODDLE II works in the field of law, the Contracts for the International Sale of Goods (CISG). The empirical results show us that DODDLE II can support a law expert in constructing domain ontologies.

2 DODDLE II: A Domain Ontology Rapid Development Environment

Figure 1 shows an overview of DODDLE II: a domain ontology rapid development environment that has the following two components: taxonomic relationship acquisition module using WordNet and non-taxonomic relationship learning module using domain-specific texts. A set of domain terms is given to DODDLE II.

The taxonomic relationship acquisition module does spell match between the input domain terms and WordNet. The spell match links these terms to WordNet. Thus the initial model from the spell match results is a hierarchically structured set of all the nodes on the path from these terms to the root of WordNet. However the initial model has unnecessary internal terms (nodes) not to contribute to keeping topological relationships among matched nodes, such as parent-child relationship and sibling relationship. So we can trim the unnecessary internal nodes from the initial model into a trimmed model, as shown in Figure 2 process. In order to refine the trimmed model, we have the following two strategies in the interaction with a user: match result analysis and trimmed result analysis that will be described later.

The non-taxonomic relationship learning module extracts the pairs of terms that should be related by some relationship from domain-specific texts, analyzing lexical co-occurrence statistics, based on WordSpace that is a multi-dimensional, real-valued vector space where the cosine of the angle be-

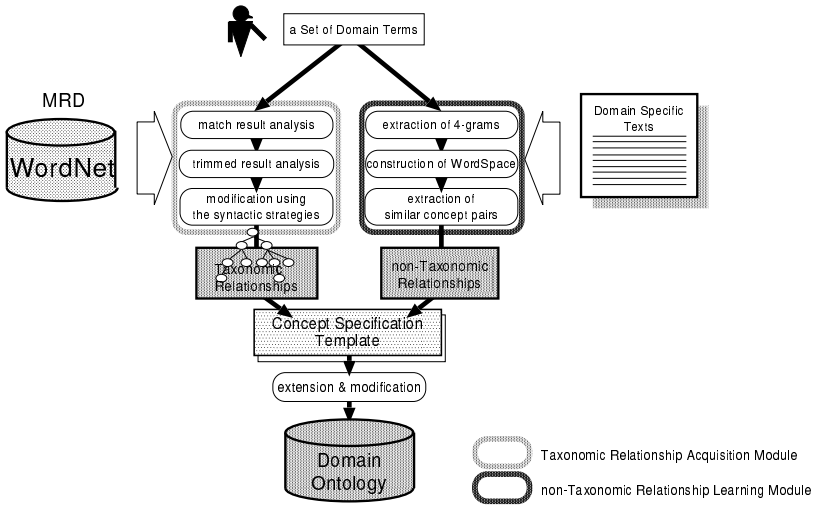

Figure 1: DODDLE II overview

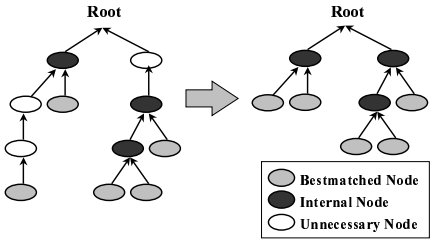

Figure 2: Trimming Process

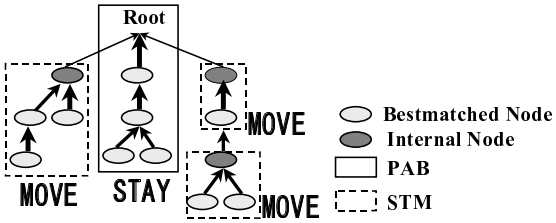

Figure 3: Match Result Analysis

tween two vectors is a continuous measure of their semantic relatedness. Thus the pairs of terms extracted from domain-specific texts are the candidates for non-taxonomic relationships. Thus putting together taxonomic and non-taxonomic relationships, we can get concept specification templates for the input domain terms, although the relationships should be identified in the interaction with a human expert.

3 Taxonomic Relationship Acquisition

After getting the trimmed model, TRA module refines it in the interaction with a domain expert, using the following two strategies: match result analysis and trimmed result analysis.

Looking at the trimmed model, it turns out that it is divided into a PAB (a PAth including only Best spell-matched nodes) and a STM (a Sub-Tree that includes best spell-matched nodes and other nodes and so should be Moved) based on the distribution of best-matched nodes. On one hand, a PAB is a path that includes only best-matched nodes that have the senses good for given domain specificity. Because all nodes have already been adjusted to the domain in PABs, PABs can stay there in the trimmed model. On the other hand, a STM is such a sub-tree that an internal node is a root and the subordinates are only best-matched nodes. Because internal nodes have not been confirmed to have the senses good for a given domain, a STM can be moved in the trimmed model. Thus DODDLE II identifies PABs and STMs in the trimmed model automatically and then supports a user in constructing a conceptual hierarchy by moving STMs. Figure 3 illustrates the above-mentioned match result analysis.

In order to refine the trimmed model, DODDLE II can use trim result analysis as well as match result analysis. Taking some sibling nodes with the same parent node, there may be many differences about the number of trimmed nodes between them and the parent node. When such a big differences comes up on a sub-tree in the trimmed model, it may be better to change the structure of the sub-tree. DODDLE II asks the user if the sub-tree should be reconstructed or not. Based on empirical analysis, the sub-trees with two or more differences may be reconstructed. Figure 4 illustrates the above-mentioned trimmed result analysis.

Finally DODDLE II completes taxonomic relationships of the input domain terms with hand-made additional modification from the user.

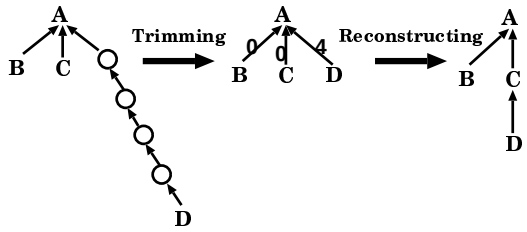

Figure 4: Trimmed Result Analysis

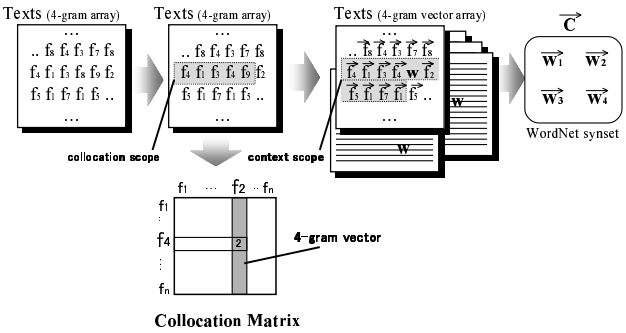

Figure 5: Construction Flow of WordSpace

4 Non-Taxonomic Relationship Learning

Non-taxonomic Relationship Learning almost comes from WordSpace[Marti and Schutze], which derives lexical co-occurrence information from a large text corpus and is a multi-dimension vector space (a set of vectors). The inner product between two word vectors works as the measure of their semantic relatedness. When two words inner product is beyond some upper bound, they are promising to have some non-taxonomic relationship between them.

4.1 Construction of WordSpace

WordSpace is constructed as shown in (Figure 5).

1. extraction of high-frequency 4-grams Since letter-by-letter co-occurrence information becomes too much and so often irrelevant, we take term-by-term co-occurrence information in four words (4-gram) as the primitive to make up co-occurrence matrix useful to represent context of a text. We take high frequency 4-grams in order to make up WordSpace.

2. construction of collocation matrix A *collocation matrix* is constructed in order to compare the context of two 4-grams. Element $a_{i,j}$ in this matrix is the number of 4-gram f_i which comes up just before 4-gram f_j (called *collocation area*). The collocation matrix counts how many other 4-grams come up before the target 4-gram. Each column of this matrix is the *4-gram vector* of the 4-gram f.

3. construction of context vectors A *context vector* represents context of a word or phrase in a text. A sum of 4-gram vectors around appearance place of a word or phrase (called *context area*) is a context vector of a word or phrase in the place.

4. construction of word vectors A word vector is a sum of context vectors at all appearance places of a word or phrase within texts, and can be expressed with the follow formula. Here, $\tau(w)$ is a vector representation of a word or phrase w, $C(w)$ is appearance places of a word or phrase w in a text, and $\varphi(f)$ is a 4-gram vector of a 4-gram f. A set of vector $\tau(w)$ is WordSpace.

5. construction of vector representations of all concepts The best matched synset of each input terms in WordNet is already specified, and a sum of the word vector contained in these synsets is set to the vector representation of a concept corresponding to a input term. The concept label is the input term.

$$\tau(w) = \sum_{i \in C(w)} \Big(\sum_{f \text{ close to } i} \varphi(f) \Big) \qquad (1)$$

4.2 Constructing and Modifying Concept Specification Templates

Vector representations of all concepts are obtained by constructing WordSpace. Similarity between concepts is obtained from inner products in all the combination of these vectors. Then, we define certain threshold for this similarity, and a concept pair with the similarity beyond it is extracted as a similar concept pair. A set of the similar concept pairs becomes concept specification templates. Both of concept pairs, which meaning is similar (with taxonomic relation), and which has something relevant each other (with non-taxonomic relation), are extracted as concept pairs with context similarity in a mass. However, by using taxonomic information from TRA module with co-occurrence information, DODDLE II distinguishes the concept pairs which hierarchically closes to each other from the other as TAXONOMY.

A user constructs a domain ontology by considering the relation with each concept pair in the concept specification templates, and deleting an unnecessary concept pair.

DODDLE II, domain ontology rapid development environment, which refer to MRD and domain-specific texts, is being implemented on Perl/Tk now. Figure 6 shows the ontology editor (left window) and the concept graph editor (right window).

5 Case Studies in the Field of Law

5.1 Learning Taxonomic Relationships

In order to evaluate how DODDLE is doing in practical fields, case studies have been done in a particular law called Contracts for the International Sale of Goods (CISG). Two lawyers joined the case studies. In the first case study, input terms are 46 legal terms from CISG Part-II. In the second case study, they are 103 terms including general terms in an example case and legal terms from CISG articles related with

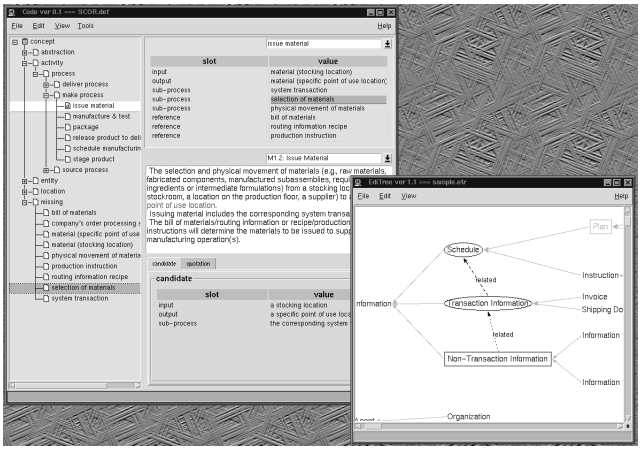

Figure 6: The Ontology Editor

the cases. One lawyer did the first case study and the other lawyer did the second.

Table 1 shows the case studies results. Figure 7 shows how much is included in final domain ontology the intermediate products at each DODDLE activity.

Generally speaking, in constructing legal ontologies, 70 % or more support comes from DODDLE. About half part of the final legal ontology results in the information extracted form WordNet. Because the two strategies just imply the part where concept drift may come up, the part generated by them has just about 30 % precision rate. Because the two strategies just take such syntactical feature as matched and trimmed results, the precision rate seems not to be so bad. In order to manage concept drift smartly, we will take into consideration the strategies with more semantic information that is not easy to come up in advance.

5.2 Learning Non-Taxonomic Relationships

We have done the case study for learning non-taxonomic relationships in the field of CISG, taking 46 legal concepts from the above-mentioned case study. A user specifies non-taxonomic relationships, taking concept specification templates from DODDLE II

Constructing WordSpace for CISG

High-frequency 4-grams have been extracted from CISG (about 10,000 words). Duplications have been removed by doing standard form conversion. Thus we have got 526 kinds of 4-grams. In order to avoid sparseness of a collocation matrix to some exten, the extraction frequency of 4-grams must be adjusted according to the scale of text corpus. As CISG is comparatively small scale, it has been set to 8 times in this case study. The collocation matrix has been constructed by counting the number of each 526 kinds 4-gram just before a 4-gram for each kind. Since 526 kinds of 4-grams have been extracted, the collocation matrix have 526 dimensions. In order to construct a context vector, we have calculated the sum of 4-gram vectors around appearance place circumference of each of 46 concepts. One article of CISG consists of about 140 4-grams. The number of 4-gram vectors in context area has been set to 60 from an experience. For each of 46 concepts, the sum of context vectors in all the appearance places of the concept in CISG has been calculated and then the vector representation of the concepts has been obtained. The set of these vectors have been used as WordSpace to extract concept pairs with context similarity.

Constructing and Modifying Concept Specification Templates

Having calculated the similarity from the inner product for the 1035 concept pairs that are all the combination of 46 concepts with the threshold of 0.9993, we have extracted 90 concept pairs and constructed concept specification templates based on them. Figure 8 illustrates the concept ”assent” specification template constructed by the user. In Figure 8, ”act” and ”proposal” have been identified as ancestor, descendant or a sibling of ”assent”, taking the concept hierarchical structure in the first case study. Thus the relationships have been distinguished as the label of TAXONOMY. As taxonomic relationships and non-taxonomic ones may come up together in the list of context similarity, it is useful to identify the taxonomic relationships by the concept hierarchical structure that has already been constructed. Given concept specification templates to a user, (s)he fills the kind of relationships in the templates, concept specifications have been done. Figure 9 illustrates the complete specification of the concept ”assent” from the concept specification template shown in Figure 8.

assent	*non-TAXONOMY?*	: **offeror**
	TAXONOMY	: **act**
	non-TAXONOMY?	: **effect**
	non-TAXONOMY?	: **offer**
	non-TAXONOMY?	: **person**
	non-TAXONOMY?	: **offeree**
	non-TAXONOMY?	: **withdrawal**
	non-TAXONOMY?	: **time**
	TAXONOMY	: **proposal**

Figure 8: The concept specification templates for “assent”

assent	AGENT	: **person**
	LEGAL-SEQUENCE	: **offer**
	ANTONYM	: **withdrawal**

Figure 9: The concept definition for ”assent” with editing the templates

5.3 Results and Evaluation

The user with legal knowledge has evaluated how much the extraction of concept pairs has been done properly. The extracted concept pairs come up in Table 2.

Figure 10 shows the trade-off between precision and recall, changing the threshold of context similarity. In getting high hit rate, coverage is small. In getting high coverage, hit rate is low. Although proper threshold exists depending on task-domains, it is hard to set it up in advance. We have not yet identified what relationship exists between two concepts. In order to do so, we need more information resources.

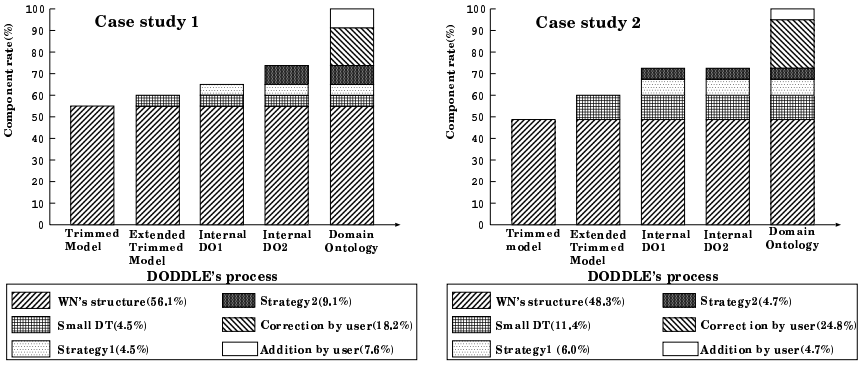

Figure 7: The Component Rate of the Final Domain Ontology

Table 1: The Case Studies Results

The number of X	The first case study	The second case study
Input terms	46	103
Small DT(Component terms)	2(6)	6(25)
Nodes matched with WordNet(Unmatched) *	42(0)	71(4)
Salient Internal Nodes(Trimmed nodes)	13(58)	27(83)
Small DT integrated into a trimmed model(Unintegrated)	2(0)	5(1)
Modification by the user(Addition)	17(5)	44(7)
Evaluation of strategy1 **	4/16(25.0%)	9/29(31.0%)
Evaluation of strategy2 **	3/10(30.0%)	4/12(33.3%)

* "Nodes matched with WordNet" is the number of input terms which have be selected proper senses in WordNet and "Unmatched" is not the case.

** The number of suggestions accepted by a user/The number of suggestions generated by DODDLE

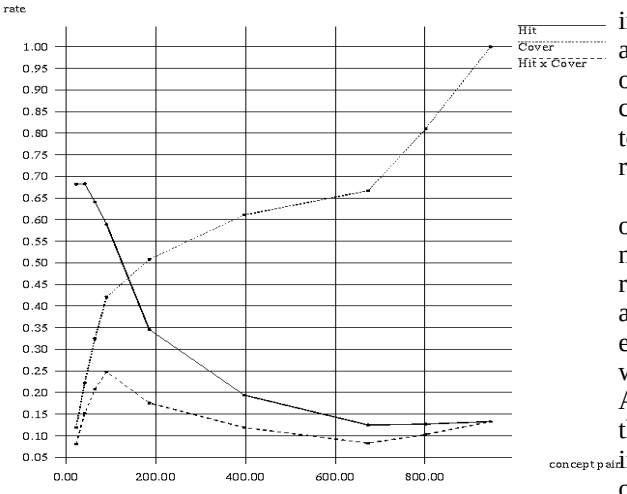

Figure 10: recall and precision

6 Related Work

On one hand, in the field of ontology learning based on natural language processing centered methods, Hahn et.al. have presented the verb-oriented methods that take the relationships of a verb and nouns modified with it, constructing concept definitions based on them (e.g. [Hahn 1998]). In [Faure and Nédellec 1999], they extract taxonomic relationships and sub-categorization Frame of verbs (SF) from technical texts using machine learning methods. They put together the nouns in two or more kinds of different SF with a same frame-name and slot-name into one concept (base class). They also build ontologies with only taxonomic relationships by doing the clustering of the base classes. Although this approach seems to be promising, the evaluation based on case studies in the real problems has not yet been done.

On the other hand, in the field of ontology learning based on data mining methods, it has been presented to discover non-taxonomic relationships using a generalized association rule algorithm by [Maedche and Staab 2000]. They explore a new metrics called RLA (Relation Learning Accuracy) to evaluate the proper abstraction level of relations, comparing with the ontology manually constructed by a human expert. Although there are similar points between their approach and the approach here, the mail goal here is to merge and exploit information resources to support the construction of domain ontologies.

7 Conclusion

Here in this paper, we have discussed how to construct a domain ontology using existing MRD and domain-specific text corpus. In order to acquire taxonomic relationships, the following two strategies have been proposed: matched result analysis and trimmed result analysis. Furthermore, in order to learn non-taxonomic relationships, the templates for concept definition has been constructed on the basis of the co-occurrence information in domain-specific text corpus, taking the construction of WordSpace. The templates work as the prototypes of concept definition. Although small-scale

Table 2: The detail of the extracted concept pairs

Threshold	Extracted concept pair	Advisable	Unknown	Improper
0.9993	90	53	14	23

case studies have been done in the field of law in order to see how DODDLE II is going in interaction with a user, we must do them in the large scale of case studies. We will consider how to take WEB content as the third information resources.

Acknowledgments

I have many thanks to Dr. Masaki Kurematsu for joining the case study and Mr. Maski Iwade for the implementations of DODDLE II.

References

[Swartout et. al. 1996] Bill Swartout, Ramesh Patil, Kevin Knight and Tom Russ: "Toward Distributed Use of Large-Scale Ontologies", Proc. of the 10th Knowledge Acquisition Workshop (KAW'96), (1996)

[Kurematsu and Yamaguchi 1997] Masaki Kurematsu and Takahira Yamaguchi: "A Legal Ontology Refinement Support Environment Using a Machine-Readable Dictionary", "*Artificial Intelligence and Law 5*", 119-137, (1997)

[Sekiuchi et. al. 1998] Rieko Sekiuchi, Chizuru Aoki, Masaki Kurematsu and Takahira Yamaguchi: "DODDLE : A Domain Ontology Rapid Development Environment", PRICAI98, (1998)

[Maedche and Staab 2000] Alexander Maedche, Steffen Staab: "Discovering Conceptual Relations from Text", ECAI2000, pp.321-325 (2000)

[Fellbaum 1998] C.Fellbaum ed: "Wordnet", The MIT Press, 1998. see also URL: http://www.cogsci.princeton.edu/~wn/

[Marti and Schutze] Marti A. Hearst, Hinrich Schutze: "Customizing a Lexicon to Better Suit a Computational Task", in Corpus Processing for Lexical Acquisition edited by Branimir Boguraev & James Pustejovsky, pp.77-96

[Hahn 1998] Udo Hahn, Klemens Schnattinger : "*Toward Text Knowledge Engineering*", AAAI98, IAAAI-98 proceedings, pp.524-531 (1998)

[Faure and Nédellec 1999] David Faure, Claire Nédellec, "Knowledge Acquisition of Predicate Argument Structures from Technical Texts Using Machine Learning: The System ASIUM", EKAW'99

[Sogano and Yamate 1993] Kazuaki Sogano, Masasi Yamate: United Nations convention on Contracts for the International Sale of Goods, Seirin-Shoin(1993)