Automatic Acquisition of Usage Information for Language Resources

Shunsuke Kozawa[†], Hitomi Tohyama[††], Kiyotaka Uchimoto[†††] and Shigeki Matsubara[††]

[†]Graduate School of Information Science, Nagoya University
[††]Information Technology Center, Nagoya University
Furo-cho, Chikusa-ku, Nagoya, 464-8601, Japan
[†††]National Institute of Information and Communications Technology
3-5 Hikari-dai, Seika-cho, Soraku-gun, Kyoto, 619-0289, Japan
{kozawa,hitomi}@el.itc.nagoya-u.ac.jp, uchimoto@nict.go.jp, matubara@nagoya-u.jp

Abstract

Recently, language resources (LRs) are becoming indispensable for linguistic research. Unfortunately, it is not easy to find their usages by searching the web even though they must be described in the Internet or academic articles. This indicates that the intrinsic value of LRs is not recognized very well. In this research, therefore, we extract a list of usage information for each LR to promote the efficient utilization of LRs. In this paper, we proposed a method for extracting a list of usage information from academic articles by using rules based on syntactic information. The rules are generated by focusing on the syntactic features that are observed in the sentences describing usage information. As a result of experiments, we achieved 72.9% in recall and 78.4% in precision for the closed test and 60.9% in recall and 72.7% in precision for the open test.

1. Introduction

In recent years, language resources (LRs) such as corpora and dictionaries have been actively used for language research. LRs are widely recognized as important and have been constructed as a research infrastructure. However, existing LRs are not fully utilized because it is not well known that they have a variety of usages. Unfortunately, it is not easy to find the usages by searching the Web even though they must be described in the Internet or academic articles. If the "usage information" such as the usages of LRs could be listed and easily referred to, the intrinsic value of LRs would be recognized and hopefully each LR would be fully utilized. For example, the usage information can be used as a query for searching for appropriate LRs. It would help users to find and efficiently use appropriate LRs if the list of usage information could be used for retrieving the catalogue information on LRs such as title, language, and samples. In the case that no appropriate language resources are found, it would become useful information for the future development of LRs because it indicates that there is a missing LR that has special needs. Furthermore, some of the usage information is not originally considered by the developer, and it would lead us to find new applications for each LR. We can expect that a list of usage information has potential to promote the effective utilization of LRs.

In this research, we extract a list of usage information for each LR to promote the efficient utilization of LRs. In this paper, we propose a method for extracting the list of usage information from academic articles by using pattern matching rules based on syntactic information. The rules are generated by focusing on the syntactic features that are observed in the sentences describing usage information. A list of usage information can be extracted by matching the rules with sentences in the articles.

2. Related Works

Information extraction from text is an active research area initiated since the MUC (Message Understanding Conference) (Grishman and Sundheim, 1996) in United States. Several information extraction tasks were organized during the MUC.

The telic role representing the typical function of the entity and the agentive role representing the origin of the entity take an important role in generative lexicon theory (Pustejovsky, 1995). For example, for the noun "book", "read" is a telic role verb and "write" is an agentive role verb. Many methods for acquiring noun-verb pairs representing telic roles or agentive roles were proposed; two methods acquiring the pairs from WordNet (Boni and Manandhar, 2002; Veale, 2003), two from corpora (Bouillon, 2002; Yamada and Baldwin, 2004), and one from the Internet (Cimiano et al., 2007). However, they are typical relationships between nouns and verbs and different from what we are aiming to extract.

There are some related works focusing on "usage information" which we attempt to extract. Inui et al. proposed a method for extracting means relations corresponding to the usage information by using Japanese cue phrase markers such as a conjunction "*tame*"(because) (Inui et al., 2005). Our proposed method uses verbs as well as conjunctions to extract a list of usage information. Torisawa proposed a method for acquiring utilization roles and preparation roles by using co-occurrence frequencies of noun, verb and postposition (Torisawa, 2005). He defined the expressions "using X" as normal manners of using X and acquired general usage of X. However, we extract all kinds of usages while he focused only on general ones because one of our purposes is to find specific usages as well as general ones for each LR. Montemagni et al. proposed a method for extracting a PURPOSE relationship which was one of semantic relations from definition in dictionaries by using syntactic information (Montemagni and Vanderwende, 1992). They mentioned that the use of syntactic information is more optimal for extracting semantic information than string patterns alone. Our extraction method also uses syntactic information. However, our target is academic articles which

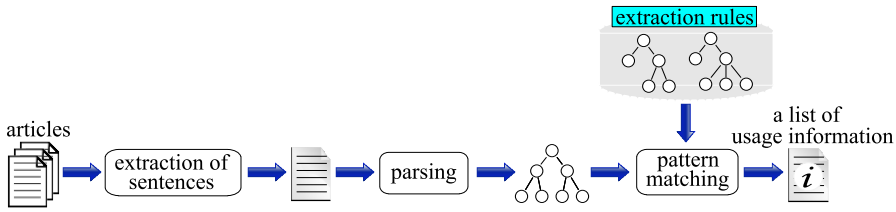

Figure 1: Processing flow.

include more varieties of description than that of their target. Therefore, no one knew how to use syntactic information appropriately for extracting usage information for LRs.

3. Analysis of Usage Information

3.1. Definition of Usage Information

Various types of description of LRs are found in available text such as academic articles and Wikipedia. For example, the following are the descriptions of WordNet which is one kind of LRs.

(1) We use WordNet for lexical lookup.

(2) We extract lexical relations from WordNet.

(3) WordNet contains semantic relationships.

(4) Resolution of pronouns can be used on top of the WordNet approach.

After investigating the contents of the sentences including titles of LRs, we found that they can be classified into mainly four types of descriptions as follows.

1. Purpose
2. Means
3. Explanation of the language resource itself
4. Miscellaneous

The first type represents purposes in use of the LR as shown in the underlined part of the example (1). The second type represents how to use the LR as shown in the underlined part of the example (2). The third type represents what is the LR is as shown in the underlined parts of the examples (3). The other sentences that are not classified into one of the three types often include information on the other LRs instead of the target LR. In this paper, we define the first and second types as usage information for the target LR.

3.2. Extraction of a List of Usage Information

The usage information can be found in both academic articles and web pages. In this paper, we chose academic articles as sources from which we extract a list of usage information because all of academic articles are not included in the Internet and more usage information can be easily found in academic articles than web pages.
We use the characteristics in describing usage information to generate extraction rules. Therefore, we focused on WordNet as a target LR and manually acquired a list of usage information from the proceedings of LREC2004 to analyze the characteristics for describing usage information. Consequently, usage information was extracted from 193 out of 214 sentences because they contained usage information. Examples of the purpose and means type are as follows. The underlined parts in the following examples represent usage information.

- Purpose
  - We use WordNet for lexical lookup.
  - The use of WordNet enables a more systematic and more detailed attachment of such marks.
  - WordNet is a valuable resource for semantic annotation.
  - The assumed baseline is the algorithm that tags the corpus according to the first WordNet sense.
- Means
  - We outline a mechanism for deriving new concepts from WordNet using metonymy.
  - Finally we assign to each noun its corresponding WordNet code.

4. Extracting a List of Usage Information by Using Syntactic Information

4.1. Overview of the Extraction Process

The flow of extracting a list of usage information is shown in Figure 1. First, we extract sentences containing the title of the target LR from academic articles. The pdftotext tool [1] can be used to convert pdf format to plain text. Next, we parse the extracted sentences using the Charniak parser (Charniak, 2000). Finally, we extract a list of usage information for a LR by applying extraction rules to the parsing result.

4.2. Extraction Rules

We generate the extraction rules by analyzing extracted usage information for WordNet from the proceedings of LREC2004 in Section3.2.
Verbs play an important role in describing usage information according to the analysis of a list of usage information. Therefore, we extract verbs and verbal phrases which are used in describing usage information and classify them into six categories by focusing on the following three points , and then generate extraction rules for each classified category. The first point is the type of verbs or verbal phrases. A copula and a general verb are discriminated. The second point is the number of objects the verb takes and the third point is the position where the title of the target LR appears in a given sentence.

[1] http://www.foolabs.com/xpdf/

- **Usage**
  The following general verbs that take an object are used and the title of the target LR appears in the object.

  use, utilize, exploit, employ, apply, leverage, etc.

  When the verb does not take any prepositions, the syntactic structure is illustrated in Figure 2. Otherwise, it is illustrated in Figure 3.

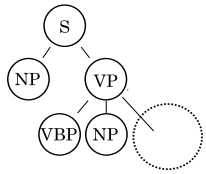

Figure 2: Usage 1.

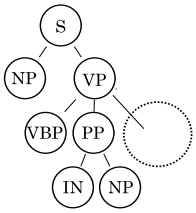

Figure 3: Usage 2.

  We extract the part circled with a dotted line as usage information if the part circled with a dotted line corresponds to one of syntactic structures as illustrated in Figure 4 through 7. Note that the preposition "IN" in Figure 4 and 5 must be one of "for", "in", "on", "as" and "towards".

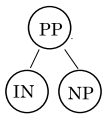

Figure 4: Purpose 1.

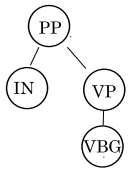

Figure 5: Purpose 2.

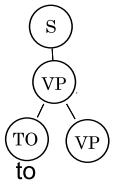

Figure 6: Purpose 3.

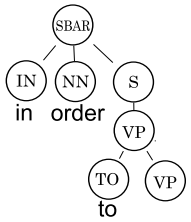

Figure 7: Purpose 4.

- **Contribution**
  The following general verbs that take objects are used and the title of the target LR appears in the subject of the given sentence.

  contribute, enable, allow, provide, help, etc.

  We extract the verb phrase circled with a dotted line as usage information if the given sentence contains one of the syntactic structures as illustrated in Figure 8 and 9.

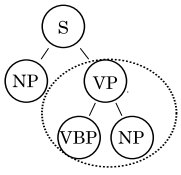

Figure 8: Contribution 1.

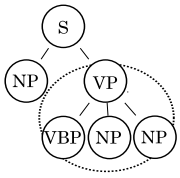

Figure 9: Contribution 2.

- **Derivation**
  The following general verbs that take two objects are used and the title of the target LR appears in the prepositional object.

  derive, obtain, extract, acquire, etc.

  We extract the verb plus noun phrase circled with a dotted line as usage information if the preposition is "from" and the given sentence contains the syntactic structure as illustrated in Figure 10.

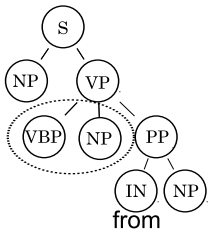

Figure 10: Derivation.

- **Linkage**
  The following general verbs that take two objects are used and the title of the target LR appears in the object of the given sentence or in the prepositional object. The following verbs are used to link or match information in LR with others. This type of pattern depends on types of LRs because these descriptions are frequently used in cases when the target LR is a conceptual dictionary such as WordNet.

  assign, match, link, merge, map, etc.

  We extract a verb phrase in the part circled with a dotted line as usage information if the given sentence contains one of the syntactic structures having prepositions as illustrated in Figure 11 through 13. If syntactic information is not used, it is difficult to determine which part should be extracted when the target verb takes two objects.

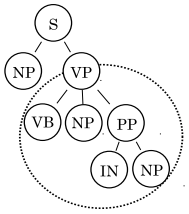

Figure 11: Linkage 1.

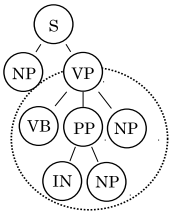

Figure 12: Linkage 2.

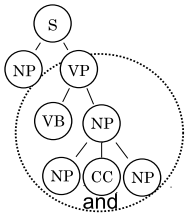

Figure 13: Linkage 3.

- **Explanation**
  Copula or the following adjectives are used and the title of the target LR appears in the subject of the given sentence.

  useful, valuable, available, helpful, etc.

  We extract the part circled with a dotted line as usage information if the given sentence has the syntactic structure as illustrated in Figure 14 or 15 and the part circled with a dotted line contains one of the syntactic structures as illustrated in Figure 4 through 6.

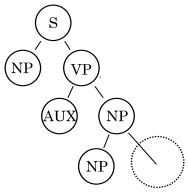

Figure 14: Explanation 1.

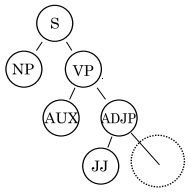

Figure 15: Explanation 2.

- **Source**
  The following verbal phrases are used.

  according to, based on, by means of, etc.

  A verb phrase or a noun phrase in the part circled with a dotted line is extracted if the given sentence contains the syntactic structure as illustrated in the verbal phrases contained in Figure 16 and 17.

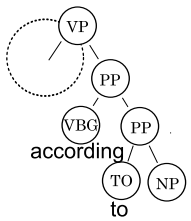

Figure 16: Source 1.

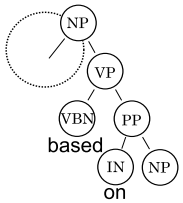

Figure 17: Source 2.

The extraction rules focusing on a general verb as mentioned above can be applied if the verb is in the active voice and present form. However, various types of descriptions using a general verb are acceptable. Therefore, we generated extraction rules taking account of gerund or past participles of a general verb.
We extract Usage and Contribution, Explanation, Source type as purpose type, and Derivation and Linkage type as means type.

5. Evaluation

We carried out the experiments to evaluate the extraction rules shown in Section 4. In this section, we tested whether our extraction rules are general by the closed and open tests. The closed test is an experiment with data which are used to generate the extraction rules while the open test is an experiment with data which are not used.

5.1. Experiments using LREC and WordNet

For a closed test, we used 214 tokens of usage information for WordNet extracted from the proceedings of LREC2004 which were used to generate the extraction rules. Furthermore, we analyzed the proceedings of LREC2006 and found that 197 tokens of usage information for WordNet. We used them for an open test.
The results of closed and open tests are shown in Table 1. We achieved 72.9% in recall and 78.4% in precision for the closed test and 60.9% in recall and 72.7% in precision for the open test. These results show that our extraction rules have a potential to extract lists of usage information for LRs.

5.2. Baseline Experiment

The feature of our proposed method is the use of syntactic information. In this section, we compared our proposed method with the baseline method that does not use syntactic information to show the advantage of syntactic information. Sentences that contain the title of the target LR and keywords such as verbs or verbal phrases were extracted with the baseline method. In this experiment, we evaluated recall (the ratio of sentences that were extracted successfully to sentences that contain usage information) and precision (the ratio of sentences that were extracted successfully to sentences that were extracted automatically).
The results of closed and open tests are shown in Table 2. They indicated that our proposed method using syntactic information has a high degree of usability.

Table 1: Results obtained from LREC using WordNet.

	LREC2004(closed)		LREC2006(open)	
	Recall(%)	Precision(%)	Recall(%)	Precision(%)
Purpose	71.3 (102/143)	80.3 (102/127)	60.9 (70/115)	70.0 (70/100)
Means	76.1 (54/71)	75.0 (54/72)	61.0 (50/82)	76.9 (50/65)
Total	72.9 (156/214)	78.4 (156/199)	60.9 (120/197)	72.7 (120/165)

Table 2: Results of Comparative Experiments.

		Recall(%)	Precision(%)	F value
LREC2004	baseline	93.3 (180/193)	28.2 (180/638)	43.3
(closed)	proposed method	77.7 (150/193)	78.1 (150/192)	77.9
LREC2006	baseline	90.2 (157/174)	27.6 (157/568)	42.3
(open)	proposed method	67.8 (118/174)	70.2 (118/168)	69.0

Table 3: Results obtained from LREC using FrameNet.

	LREC2004(closed)		LREC2006(open)	
	Recall(%)	Precision(%)	Recall(%)	Precision(%)
Purpose	75.0 (6/8)	85.7 (6/7)	78.9 (15/19)	78.9 (15/19)
Means	50.0 (3/6)	100 (3/3)	69.2 (9/13)	90.0 (9/10)
Total	64.3 (9/14)	100 (9/9)	75.0 (24/32)	82.8 (24/29)

Table 4: Results obtained from LREC using Penn Treebank.

	LREC2004(closed)		LREC2006(open)	
	Recall(%)	Precision(%)	Recall(%)	Precision(%)
Purpose	83.3 (10/12)	83.3 (10/12)	38.5 (5/13)	55.6 (5/9)
Means	65.0 (13/20)	76.5 (13/17)	72.7 (8/11)	72.7 (8/11)
Total	67.6 (23/34)	79.3 (23/29)	54.2 (13/24)	65.0 (13/20)

5.3. Experiments Using Other LRs

Usage information for other LRs can be described in a different way from that of WordNet. Therefore, the extraction rules were applied to other LRs to know whether the rules are general enough.

5.3.1. Applying Rules to Same Types of LRs

Linkage type depends on types of LRs. Therefore, we applied extraction rules to FrameNet which is the same type of LRs as WordNet.
The results of the closed and open tests are shown in Table 3. The comparative results with WordNet were obtained for FrameNet without major changes to the extraction rules. These results show that we can extract usage information for LRs which are the same type of LRs as WordNet by using extraction rules without major changes.

5.3.2. Applying Rules to Different Types of LRs

The extraction rules were applied to Penn Treebank which is different type of LRs than WordNet. First, we carried out the open test using a list of usage information for Penn

Table 5: Results obtained from SLP using WordNet.

	open		closed		open	
	Recall(%)	Precision(%)	Recall(%)	Precision(%)	Recall(%)	Precision(%)
EUROSPEECH	60.0 (3/5)	75.0 (3/4)	60.0 (3/5)	75.0 (3/4)	————	————
ICASSP	0 (0/1)	0 (0/2)	100 (1/1)	100 (1/1)	————	————
ICSLP	33.3 (1/3)	25.0 (1/4)	33.3 (1/3)	33.3 (1/3)	47.6 (10/21)	81.8 (9/11)

Treebank extracted from the proceedings of LREC2004. Consequently, the precision was 75.0%. However the recall was 44.1% because we could not extract means type including Linkage type. Therefore, we investigated the proceedings of LREC2004 to find characteristic expressions to Penn Treebank, and acquired the following verbs.

convert, translate, transform, parse, train

We added these verbs to the extraction rules and performed experiments.
The results of the closed and open tests are shown in Table 5.3.1.. It is possible to apply to different types of LRs by acquiring expressions which depend on the type of LRs, because usage information described by using expressions which depend on Penn Treebank were extracted in the open test.

5.4. Experiments Using Other Academic Articles

We applied the extraction rules to other academic articles and extracted a list of usage information for WordNet.

5.4.1. Applying Rules to the Academic Articles in the Field of Spoken Language Processing

The extraction rules were applied to the proceedings of EUROSPEECH, ICASSP and ICSLP in the field of spoken language processing (SLP) to know whether our extraction rules are applicable in other fields. We investigated the two proceedings of each conference. However, there is only one proceeding which contains a list of usage information for WordNet in the proceedings of EUROSPEECH and ICASSP. Therefore, we carried out the open and closed tests using same proceedings. In addition, we carried out the open and closed tests using the proceedings of ICSLP2004 and the open test using the proceedings of ICSLP2006.
The results of the closed and open tests are shown in Table 5. Though the number of extracted lists of usage information are small, these results show that the extraction rules are applicable to academic articles in other fields with minor changes.

5.4.2. Applying Rules to the Proceedings of ACL

We applied the extraction rules to the proceedings of ACL and extracted a list of usage information for WordNet. The results of the closed and open tests are shown in Table 6. We achieved 69.4% in recall and 76.1% in precision for the closed test and 59.0% in recall and 68.1% in precision for the open test. The results obtained from ACL was lower than LREC because the parsing errors appeared more frequently. However, we believe that extraction rules were more general because difference between the results of the closed and open tests decreased. In addition, we can improve the results of the open test without major changes.

Table 6: Results obtained from ACL using WordNet.

	ACL2004(closed)		ACL2005(open)	
	Recall(%)	Precision(%)	Recall(%)	Precision(%)
Purpose	71.7 (91/127)	71.1 (91/128)	66.0 (64/97)	66.0 (64/97)
Means	65.8 (52/79)	86.7 (52/60)	40.5 (15/37)	78.9 (15/19)
Total	69.4 (143/206)	76.1 (143/188)	59.0 (79/134)	68.1 (79/116)

5.5. Examples of Extracted Usage Information

Examples of successfully extracted lists of usage information are shown below. The following are the examples of purpose type of usage information.

- for NLP
- for word sense disambiguation
- for query expansion
- to cluster its senses

We could acquire practical purposes such as disambiguation, clustering and query expansion as shown in the second through fourth examples although some were general purposes as shown in the first example.
The following are the examples of means type of usage information. We could acquire practical means for disambiguation and clustering.

- extract a lexical expression
- assign WordNet senses to cluster labels

5.6. Discussion

In this paper, sophistication of extraction rules were performed by application to different LRs and academic articles. Sophistication in extraction rules contained the following. The first action has the largest effect on extraction rules while the third action has the smallest effect.

1. Addition of new extraction rules
2. Revision of the extraction rules
3. Addition of characteristic verbs or expressions to the extraction rules

In the sophistication process, we found that the third action was the most frequent one while the first action was rare. That is, major changes of the rules were rare and the most major action was slight modification. The second and third actions will be required to sophisticate the extraction rules. However, we believe that our extraction rules would

become general enough for various LRs because the number of actions required to sophisticate the rules decreased during each step of applying the rules to the three types of LRs or other academic articles.

In this paper, lists of usage information were extracted from sentences containing the title of the target LR. Therefore, sentences that do not contain the title of the target LR were ignored even if they contain usage information. In these sentences, the pronoun or the explanation of the target LR are often used instead of the title of the LR. It would be possible to extract usage information from such sentences in the following way. For the sentences having a pronoun, usage information can be extracted by taking account of anaphoric relations. For the sentences having the explanation of the target LRs, usage information can be extracted in the following steps. First, explanations of LRs are obtained by using rules for extracting explanation type of knowledge. Then, the LR title is detected for each explanation by using lists of the pair of LR title and its explanation which are extracted beforehand. Finally, our method is applied to sentences if the detected LR title corresponds to that of the target LR.

We investigated the difference between lists of usage information for WordNet extracted from LREC2004 and LREC2006 and found that about 30 percent of lists extracted from LREC2006 were not found in those extracted from LREC2004. For example, the following expressions were newly found in LREC2006.

- in the biomedical domain
- improve interoperability, user-friendliness and usability of both lexical resources
- combined geographical databases with WordNet

This indicates that new usages are being created for WordNet. It also indicates that new usage information would be extracted from other collections of articles.

6. Conclusion

In this paper, we proposed a method for extracting a list of usage information for each LR from academic articles by using rules based on syntactic features to promote the effective utilization of LRs. The rules are generated by focusing on the syntactic features that are observed in the sentences describing usage information. Experimental results show that our extraction rules are applicable to other LRs or academic articles without major changes. We believe that the acquisition of usage information for LRs is at practical level.

Our future works will focus on sophistication of extraction rules and applying our method to the Internet. In this paper, we targeted academic articles. However, we would like to apply our methods to the Internet because it has the potential to contain new types of usage information. The extraction rules should be sophisticated enough to apply them to the Internet because there are some differences between descriptions in academic articles and the web pages. In addition, we would like to research whether lists of usage information have an effect on the search for LRs by applying extracted usage information to the metadata database of LRs named SHACHI (Tohyama et al., 2008) in the near future.

7. References

Bouillon, P., Claveau, V., Fabre, C. & Sebillot, P. (2002). *Acquisition of Qualia Elements from Corpora - Evaluation of a Symbolic Learning Method*. In Proceedings of the 3rd International Conference on Language Resources and Evaluation, pp. 208-215.

Charniak, E. (2000). *A Maximum-entropy-inspired Parse*. In Proceedings of the North American chapter of the Association for Computational Linguistics, pp.132-139.

Cimiano, P. & Wenderoth, J. (2007). *Automatic Acquisition of Ranked Qualia Structures from the Web*. In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, pp. 888-895.

De Boni, M. & Manandhar, S. (2002). *Automated Discovery of Telic Relations for WordNet*. In Proceedings of the 1st International WordNet Conference.

Grishman, R. & Sundheim B. (1996). *Message Understanding Conference - 6: A Bried History*. In Proceedings of the 16th International Conference on Computational Linguistics, pp. 466-471.

Inui, T., Inui, K. & Matsumoto, Y. (2005). *Acquiring Causal Knowledge from Text Using Connective Marker Tame*. ACM Transactions on Asian Language Information Processing, 4(4), pp. 435-474.

Montemagni, S. & Vanderwende, L. (1992). *Structural Patterns vs. String Patterns for Extracting Semantic Information from Dictionaries*. In Proceedings of the 14th International Conference on Computational Linguistics, pp. 546-552.

Pustejovsky, J. (1995). *The Generative Lexicon*. MIT Press.

Tohyama, H., Kozawa, S., Uchimoto, K., Matsubara, S. & Isahara, H. (2008). *SHACHI: A Large Scale Metadata Database of Language Resources*. In Proceedings of the 1st International Conference on Global Interoperability for Language resources, pp. 205-212.

Torisawa, K. (2005). *Automatic Acquisition of Expressions Representing Preparation and Utilization of an Object*. In Proceedings of the Recent Advances in Natural Language Processing, pp. 556-560.

Veale, T. (2003). *Qualia Extraction and Creative Metaphor in WordNet*. In Proceedings of the 18th International Joint Conference on Artificial Intelligence.

Yamada, I. & Baldwin, T. (2004). *Automatic Discovery of Telic and Agentive Roles from Corpus Data*. In Proceedings of the 18th Pacific Asia Conference on Language, Information and Computation, pp. 115-126.