Collection of Usage Information for Language Resources from Academic Articles

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
4-2-1 Nukui-Kitamachi, Koganei, Tokyo, 184-8795, 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 researches. However, existing LRs are often not fully utilized because their variety of usage is not well known, indicating that their intrinsic value is not recognized very well either. Regarding this issue, lists of usage information might improve LR searches and lead to their efficient use. In this research, therefore, we collect a list of usage information for each LR from academic articles to promote the efficient utilization of LRs. This paper proposes to construct a text corpus annotated with usage information (UI corpus). In particular, we automatically extract sentences containing LR names from academic articles. Then, the extracted sentences are annotated with usage information by two annotators in a cascaded manner. We show that the UI corpus contributes to efficient LR searches by combining the UI corpus with a metadata database of LRs and comparing the number of LRs retrieved with and without the UI corpus.

1. Introduction

In recent years, such language resources (LRs) as corpora and dictionaries are being widely used for research in the fields of linguistics, natural language processing, and spoken language processing, reflecting the recognition that objectively analyzing linguistic behavior based on actual examples is important. Therefore, since the importance of LRs is widely recognized, they have been constructed as a research infrastructure and are becoming indispensable for linguistic research. However, existing LRs are not fully utilized. Even though metadata search services for LR archives (Hughes and Kamat, 2005) and web services for LRs (Dalli et al., 2004; Biemann et al., 2004; Quasthoff et al., 2006; Ishida et al., 2008) have become available, it has not been enough for users to efficiently find and use LRs suitable for their own purposes so far.

If there exists a system which could give us a list of LRs that can be answers to the questions such as "Which LRs can be used for developing a syntactic parser?" and "Which LRs can be used for developing a Chinese-English machine translation system?," it would help users efficiently find appropriate LRs.

The information satisfying these demands is sometimes described as usages of individual LRs on their official home pages. The metadata database of LRs named SHACHI (Tohyama et al., 2008) is managing and providing it in an integrated fashion by collecting and listing it as "usage information" for LRs. SHACHI contains metadata on approximately 2,400 LRs. However, the number of lists of usage information registered in SHACHI is only about 900 LRs since the usage information is not usually described on the official home while it is often described in academic articles. For instance, the following sentence found in the proceedings of ACL2006 shows that Roget's Thesaurus is useful for word sense disambiguation, although usage information is not announced on the web page of Roget's Thesaurus[1].

- He also employed *Roget's Thesaurus* in 100 words of window to implement **WSD**.

Therefore, we could more easily find LRs suitable for our own purposes by collecting lists of usage information for LRs from academic articles and integrating them with metadata contained in SHACHI. Although the method for automatically extracting the lists was proposed (Kozawa et al., 2008), the variation of the extracted usage information was limited since their extraction rules were based on the analysis of small lists of usage information. This issue would be addressed by collecting large lists of usage information and then analyzing them to expand the extraction rules. Therefore, in this paper, we propose to construct a text corpus annotated with usage information (UI corpus).

This paper is organized as follows. In sections 2, we introduce the design of UI corpus. Then, we construct the UI corpus by extracting sentences containing LR names from academic articles and annotating them with usage information in section 3. In sections 4 and 5, we provide statistics of the UI corpus and analytical results of usage information contained in the UI corpus. We show that the UI corpus contributes to efficient LR searches by combining the UI corpus with a metadata database of LRs and comparing the number of LRs retrieved with and without the UI corpus in section 6. Finally, in section 7, we describe the summary of this paper and the future work.

2. Design of the UI Corpus

2.1. Data Collection

It is unrealistic to collect all sentences and annotate them because only small number of sentences in an article include usage information for LRs. In this issue, Kozawa et

[1]http://poets.notredame.ac.jp/Roget/about.html

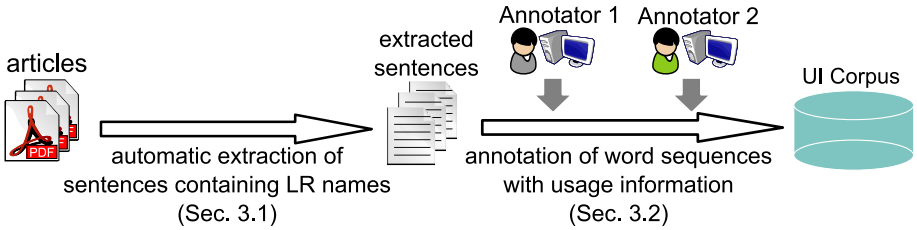

Figure 1: Flow of the UI corpus construction.

al. reported that most of the instances of usage information for LRs are found in the sentences containing LR names (Kozawa et al., 2008). Therefore, in this research, we collect sentences having LR names from academic articles to build the UI corpus.

2.2. Annotation Policy

The collected sentences are annotated with the following information: (A), (B) and (C) are provided for each sentence. (D) and (E) are provided for word sequences. (A) through (D) are automatically provided when sentences have been collected from academic articles.

(A) sentence ID

(B) the title of the proceeding

(C) article ID

(D) LR name
Word sequences matched with LR names are annotated with LR tags as the following example:

- For comparison, ⟨LR⟩Penn Treebank⟨/LR⟩ contains over 2400 (much shorter) WSJ articles.

Note that the LR tags with which word sequences are automatically annotated need to be examined, since homographs of the LR names (ex. the names of projects and associations) are sometimes erroneously annotated as LR names. For example, 'Penn Treebank' is often used as an LR name and it is sometimes used as a project name in the different context. Therefore, it is difficult to discriminate proper LR names from others. Then, if inappropriate word sequences are annotated with LR tags, they are manually eliminated.

If word sequences matched with two or more LRs have a coordinate structure as the following example, "Chinese" and "English Propbanks" are annotated with LR tags. This is because we distinguish between two or more LR names (e.g. Chinese and English PropBanks) and an LR name which have a coordinate structure (e.g. Cobuild Concordance and Collocations Sampler).

- The functional tags for ⟨LR⟩Chinese⟨/LR⟩ and ⟨LR⟩English PropBanks⟨/LR⟩ are to a large extent similar.

(E) usage information
Each word sequence matched with usage information for a certain LR is annotated with UI tags. Word sequences are annotated with usage information by referring only to a given sentence without adjacent sentences in order to reduce the labor costs of annotators.

In our research, we assume that usage information A for LR X can be paraphrased as "X is used for A." The followings are examples of usage information for WordNet.

- We use ⟨LR⟩WordNet⟨/LR⟩ for ⟨UI⟩lexical lookup⟨/UI⟩.
- ⟨LR⟩WordNet⟨/LR⟩ ⟨UI⟩specifies relationships among the meanings of words⟨/UI⟩.
- It uses the content of ⟨LR⟩WordNet⟨/LR⟩ to ⟨UI⟩measure the similarity or relatedness between the senses of a target word and its surrounding words⟨/UI⟩ .

Note that since usage information indicates specific events, such vague expressions as "our proposed method" and "this purpose" are not our target. We also ignore expressions that can be represented by "X is used for X" and those that represent updating, expansion, or modification of LR X, as shown in the following example:

- We applied an automatic mapping from ⟨LR⟩WordNet 1.6⟨/LR⟩ to ⟨LR⟩WordNet 1.7.1⟨/LR⟩ synset labels.

3. Construction of the UI Corpus

This section describes a method for constructing the UI corpus. Figure 1 shows the flow of the corpus construction. First, sentences containing LR names are automatically extracted from academic articles. Then, the extracted sentences are annotated by two annotators in a cascaded manner.

3.1. Automatic Extraction of Sentences Containing LR Names

First, we converted academic articles to plain texts using the Xpdf[2]. We used 2,971 articles which are contained in the proceedings of ACL from 2000 to 2006, LREC2004 and LREC2006 because LRs were often used in the field of computational linguistics. Next, we extracted sentences

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

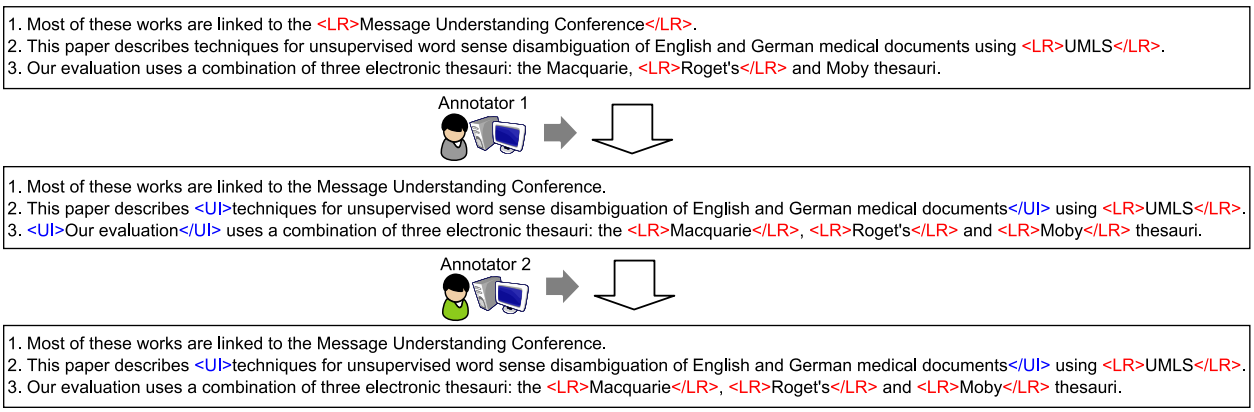

Figure 2: Flow of corpus annotation.

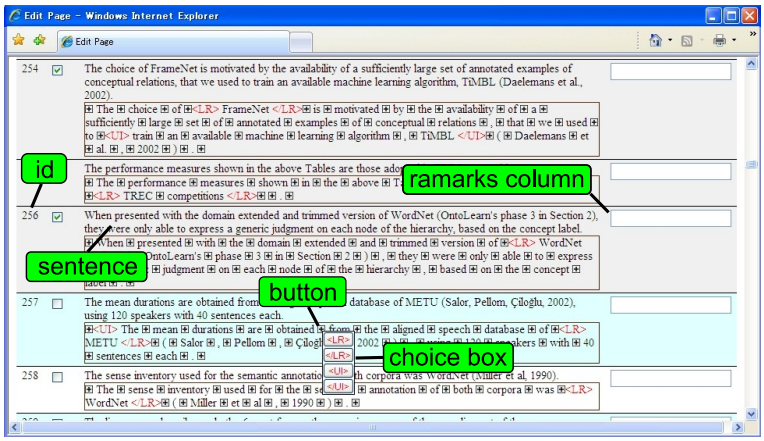

Figure 3: Web-based GUI for supporting annotation.

containing the LR names from the articles. As for the LR names, approximately 2,400 LRs registered in SHACHI (Tohyama et al., 2008) were used. Consequently, 10,959 sentences were extracted from 1,848 articles and word sequences matched with the LR names are automatically annotated with the LR tags.

3.2. Annotation of Word Sequences with Usage Information

Two annotators were involved in the annotation. One of the annotators (Annotator 1) had an experience in collecting metadata in SHACHI, while the major of the others (Annotator 2) was computational linguistics. It was difficult for Annotator 1 to annotate usage information if a given sentence contains technical terms in the field of computational linguistics, since Annotator 1 was unfamiliar with computational linguistics. Therefore, Annotator 1 annotated sentences at first, and then Annotator 2 annotated the same sentences to recover the annotation errors produced by Annotator 1.

Examples of corpus annotation are shown in Figure 2. First, the following actions were done by Annotator 1:

- Annotating word sequences representing LR names with the LR tags
- Annotating word sequences representing the usage information for LRs with the UI tags

Next, Annotator 2 judged whether the UI tags provided by Annotator 1 were correct and modified them if they were inappropriate.

For annotation, the annotators used the Web-based GUI as shown in Figure 3. They could make annotation by selecting an appropriate tag from the choice box appeared by clicking the button shown in Figure 3.

Item	Number
articles	1533
sentences	8135
LR tags	10504
sentences containing usage information	1110
UI tags	1183

Table 1: Size of the UI corpus.

3.3. Corpus Size

The size of the annotated corpus is shown in Table 3.2.. In the process of annotation, sentences which did not contain any LR names were removed from the corpus. Therefore, the corpus consists of a set of sentences containing LR

Rank	LR name	# of LR tags
1	WordNet	2356
2	Penn Treebank	745
3	FrameNet	437
4	British National Corpus	433
5	PropBank	419
6	SemCor	209
7	VerbNet	191
8	TREC Collection	166
9	MeSH	147
10	EuroWordNet	144

Table 2: Frequently appearing LRs.

names extracted from academic articles.

4. Statistics of the UI Corpus

This section shows the frequently used LRs and the difference between the LRs tagged with the UI tags in the UI corpus and the LRs whose usage information were registered in SHACHI by comparing the statistics of the UI corpus with that of SHACHI.

We semi-automatically assigned each LR tag with an LR id used in SHACHI and counted the number of LRs appearing in our corpus. Consequently, 882 LRs were found. Then, we investigated the breakdown of the LR tags. We found that the most frequent LR was WordNet. One of the reasons is that WordNet is frequently used as a lexical database. Another reason is that it has been translated into various languages by the initiative of Global WordNet Association[3].

Out of 882 LRs, 365 were tagged with the UI tags. We investigated whether the usage information for the LRs was registered in SHACHI or not, and found that usage information for 305 LRs was not registered in SHACHI. This shows that usage information for LRs newly extracted from academic articles were almost double of that originally registered in SHACHI. We expect that the more usage information could be extracted if we used more variety of academic articles and it would help users efficiently find and use LRs suitable for their own purposes by registering lists of usage information for finding more LRs than those obtained only with SHACHI.

5. Analysis of Usage Information

This section shows the number of types of usage information and LRs used in various fields.

Lists of usage information were analyzed to know how many types of usage information were collected. Then, we manually classified them into 51 classes (see Table 4.) in the fields of computational linguistics and spoken language processing, which are session names appeared twice or more in the proceedings of ACL from 2000 to 2006 or ICSLP 2000, 2002, 2004 and 2006, by referring to the articles containing target usage information. Note that each usage information is classified into one or more classes.

[3]http://www.globalwordnet.org/

Acoustic Modeling	Question Answering
Applications	Robust ASR
Asian Language Processing	Segmentation
Chunking	Semantics
Coreference	Speaker Recognition
Corpora	Speaker Segmentation
Dialect Recognition	Speech Analysis
Dialogue	Speech Coding
Discourse	Speech Enhancement
Generation	Speech Features
Grammars	Speech Perception
Information Extraction	Speech Processing
Information Retrieval	Speech Production
Language Acquisition	Speech Synthesis
Language Modeling	Speech Translation
Language Recognition	Spoken Language Processing
Large Vocabulary Speech Recognition	Spoken Language Resource
Lexical Semantics	Spoken Language Understanding
Linguistics	Statistical Machine Translation
Machine Translation	Statistical Parsing
Morphology	Summarization
Named Entity Recognition	Syntax
Parsing	Tagging
Phonetics	Text Categorization
Phonology	Word Sense Disambiguation
Prosody	

Table 3: Classes.

Rank	LR name	# of classes
1	Penn Treebank	24
2	WordNet	23
3	British National Corpus	21
4	Reuters Corpus	9
5	EuroWordNet	8
	FrameNet	8
	UMLS	8
	Chinese Treebank	8
9	Spoken Dutch Corpus	7
	TREC Collection	7
	Brown Corpus	7
	TIGER Corpus	7

Table 5: Versatile LRs.

Classification results are shown in Table 4.. Column 2 and 3 represent the number of UI tags classified into each class and the number of articles containing the UI tags, respectively. The number of LRs tagged with the UI tags are shown in column 4. In column 5, frequently used LRs in each fields are represented and parenthetical figure denotes the number of articles using the LRs. Large lists of usage information in the fields of "lexical semantics" and "word sense disambiguation" were collected since WordNet was frequently used. Out of 51 classes, 39 have one or more UI tags. This shows that the UI corpus contains various usage information.

We investigated the number of classes to which UI tags were classified for each LR to find LRs used in various fields. Results of the investigation are shown in Table 5.. We found that Penn Treebank is the most widely used LR. In addition, British National Corpus is also widely used although the frequency of British National Corpus is lower than those of WordNet and Penn Treebank.

class	# of UI tags	# of articles	# of LRs	frequently used LR
Lexical Semantics	188	110	35	WordNet(82), UMLS(4), EuroWordNet(4)
Word Sense Disambiguation	182	88	42	WordNet(61), Semcor(8), Roget's(5), Longman Dictionary(5)
Semantics	147	81	38	WordNet(27), FrameNet(18), PropBank(7)
Corpora	130	66	51	WordNet(28), Penn Treebank(4), VerbNet(4), British National Corpus(4)
Information Extraction	104	71	43	WordNet(22), British National Corpus(6), ACE(5), GENIA(4)
Parsing	66	44	30	Penn Treebank(22), CCGbank(3)
Tagging	59	40	29	Penn Treebank(11), GENIA(4), CGN(4)
Question Answering	48	26	13	WordNet(15), TREC(5), Wikipedia(2), Extended WordNet(2)
Machine Translation	44	32	31	WordNet(7), Czech Treebank(2), British National Corpus(2)
Asian Language Processing	36	30	24	Mainichi newspapers(3), Sinica Treebank(2), EDR(2)
Grammars	30	22	8	Penn Treebank(17), British National Corpus(2)
Information Retrieval	29	22	12	WordNet(9), TREC(4), NTCIR(2), EuroWordNet(2)
Coreference	29	18	15	WordNet(4), British National Corpus(2), Reuters corpus(2), ACE(2)
Named Entity Recognition	25	19	18	MUC(4), GENIA(3), UMLS(2), WordNet(2), Reuters corpus(2)
Statistical Parsing	25	15	7	Penn Treebank(10), CCGbank(2)
Syntax	18	15	10	Penn Treebank(9)
Spoken Language Processing	17	11	12	Spoken Dutch Corpus(2), Penn Treebank(2)
Dialogue	17	11	9	ICSI Meeting corpus(2), British National Corpus(2)
Chunking	16	9	8	Penn Treebank(5)
Summarization	14	11	10	Mainichi newspapers(3), English Broadcast News corpus(2)
Generation	14	9	9	WordNet(3)
Text Categorization	12	12	10	Reuters-21578 corpus(3), WordNet(2), Gigaword corpus(2)
Morphology	11	9	8	Arabic Treebank(3), Chinese Penn Treebank(2)
Robust ASR	10	8	9	SpeechDat(1), LC-Star(1), ATIS corpus(1)
Language Modeling	10	8	5	British National Corpus(3), Yomiuri Newspapers(2)
Language Acquisition	9	7	5	WordNet(3)
Statistical Machine Translation	9	6	8	WordNet(2)
Prosody	9	6	8	Spoken Dutch Corpus(1), TDT(1), TTS evaluation corpus(1)
Segmentation	9	4	8	British National Corpus(1), Penn Treebank(1), ECI corpus(1)
Phonetics	8	6	5	Spoken Dutch Corpus(2), CELEX(2)
Linguistics	8	5	5	WordNet(2)
Phonology	6	5	3	CELEX(2), Penn Treebank(2)
Discourse	4	3	3	British National Corpus(2)
Large Vocabulary Speech Recognition	3	3	3	SpeechDat(1), Slovenian broadcast news speech database(1)
Applications	3	3	3	WordNet(1), Brown Corpus(1), Mainichi Daily News(1)
Speech Processing	3	2	2	Switchborad(1), RT corpus(1)
Speech Synthesis	2	2	2	LC-Star(1), METU Turkish Corpus(1)
Acoustic Modeling	2	1	1	NIST Corpus(1)
Speech Analysis	1	1	1	Czech National Corpus(1)

Table 4: Classification results of usage information.

6. Contribution of the UI Corpus

We compared the number of LRs retrieved with and without usage information in the UI corpus. In the experiments, we used keywords as queries and got a list of LRs whose usage information registered in SHACHI or in the UI corpus contained the keywords. As queries for the LR search, we used 40 keywords in the "Topics of Interest" appearing in the paper submission page of ACL2008.

The experimental results are shown in Table 6.. The number of LRs retrieved using usage information in both SHACHI and the UI corpus increased for 15 keywords. This indicates that lists of usage information in the UI corpus contribute to efficient LR searches.

We are planning to train the model for extracting usage information for LRs by using our corpus to improve the performance of automatic usage information extraction and extract usage information from various articles. Then, we expect that more various LRs can be found.

7. Conclusion

In this paper, we described how to construct the UI corpus to efficiently find and use appropriate LRs. First, we automatically extracted sentences containing LR names from academic articles. Then, two annotators tagged the extracted sentences with usage information. We showed that the UI corpus contributes to efficient LR searches by combining the UI corpus with a metadata database of LRs.

In the near future, we will provide an LR search service to promote the efficient use of LRs by integrating usage information with a metadata database of LRs called SHACHI (Tohyama et al., 2008).

Acknowledgments

This research was supported in part by The Hori Information Science Promotion Foundation.

8. References

Christian Biemann, Stefan Bordag, Uwe Quasthoff, and Christian Wolff. 2004. Web services for language resources and language technology applications. In *Proceedings of 4th International Conference on Language Resources and Evaluation*.

Angelo Dalli, Valentin Tablan, Kalina Bontcheva, Yorick Wilks, Dan Broeder, Hennie Brugman, and Peter Wittenburg. 2004. Web services architecture for language resources. In *Proceedings of 4th International Conference on Language Resources and Evaluation*.

Baden Hughes and Amol Kamat. 2005. A metadata search engine for digital language archives. *DLib Magazine*, 11(2):6.

Toru Ishida, Akiyo Nadamoto, Yohei Murakami, Rieko Inaba, Tomohiro Shigenobu, Shigeo Matsubara, Hiromitsu Hattori, Yoko Kubota, Takao Nakaguchi, and Eri Tsunokawa. 2008. A non-profit operation model for the

Keywords	SHACHI	SHACHI+UI corpus
dialogue	8	12
embodied conversational agents	0	0
language-enhanced platforms	0	0
information retrieval	52	54
text data mining	0	0
information extraction	11	11
filtering	0	2
recommendation	0	0
question answering	0	3
topic classification	0	0
text classification	3	3
sentiment analysis	0	0
attribute analysis	0	0
genre analysis	0	0
language generation	1	1
summarization	8	14
machine translation	55	57
language identification	21	21
multimodal processing	0	0
speech recognition	211	228
speech generation	1	1
speech synthesis	32	34
phonology	0	0
POS tagging	1	7
syntax	0	7
parsing	11	19
grammar induction	0	0
mathematical linguistics	0	0
formal grammar	0	0
semantics	1	5
textual entailment	0	0
paraphrasing	0	2
word sense disambiguation	0	11
discourse	10	14
pragmatics	0	0
statistical and machine learning	0	0
language modeling	25	25
lexical acquisition	0	0
knowledge acquisition	0	0
development of language resources	0	0

Table 6: Results of verification experiment using the UI corpus.

language grid. In *Proceedings of the 1st International Conference on Global Interoperability for Language resources*, pages 114–121.

Shunsuke Kozawa, Hitomi Tohyama, Kiyotaka Uchimoto, and Shigeki Matsubara. 2008. Automatic acquisition of usage information for language resources. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*.

Uwe Quasthoff, Matthias Richter, and Christian Biemann. 2006. Corpus portal for search in monolingual corpora. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*.

Hitomi Tohyama, Shunsuke Kozawa, Kiyotaka Uchimoto, Shigeki Matsubara, and Hitoshi Isahara. 2008. Construction of a metadata database for efficient development and use of language resources. In *Proceedings of 6th International Conference on Language Resources and Evaluation*.