Utilization of Usage Information for Language Resource Searches

Shunsuke Kozawa, Hitomi Tohyama
Nagoya University
Furo-cho, Chikusa-ku,
Nagoya, 464-8601, Japan
{kozawa,hitomi}@
el.itc.nagoya-u.ac.jp

Kiyotaka Uchimoto
National Institute of Information
and Communications Technology
4-2-1 Nukui-Kitamachi,
Koganei, Tokyo, 184-8795, Japan
uchimoto@nict.go.jp

Shigeki Matsubara
Nagoya University
matubara
@nagoya-u.jp

Abstract

Recently, language resources (LRs) are becoming indispensable for linguistic research. 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 paper, we show that lists of usage information for each LR contribute to efficient LR searches. We combine the varieties of automatically extracted usage information with a metadata database of LRs. Then we compare the efficiency of LR searches with and without usage information.

1 Introduction

In recent years, such language resources (LRs) as corpora and dictionaries are being widely used for linguistic research. Therefore, since the importance of LRs is widely recognized, they have been constructed as a research infrastructure and are becoming indispensable for 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 (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 such "usage information" as the usage of LRs could be listed and easily referred to, their intrinsic value might be recognized and perhaps each LR would be fully utilized. In our research, we assume that usage information A for LR X can satisfy the relation "X is used for A." For example, usage information for WordNet is represented by such expressions as "lexical lookup." If a list of usage information could be used for retrieving LRs suitable for our own purposes, it would help us efficiently find and use appropriate LRs.

The usages of individual LRs are often announced on official home pages and they could be provided in an integrated fashion by collecting and listing them (Tohyama et al., 2008a). However, the usages announced on official home pages are just usages considered by the developer. In contrast to their usages, there exists information about user's experience and knowledge. To offer these user's usages is more effective in the LR search than ones considered by developer.

In this paper, we show that lists of automatically extracted usage information for each LR contribute to efficient LR searches. In particular, we combine lists of usage information by a method based on syntactic information with a metadata database of LRs. Then we compare the number of LRs retrieved with and without usage information.

2 Availability of Usage Information for LR Searches

Usage information for LRs are announced on official home pages. For instance, 6 instances of usage information (linguistics, information retrieval, word sense disambiguation and so on) are announced on the web page of WordNet[1], and 3 instances of usage information (natural language processing, information retrieval and machine learning) are announced on the web page of Rueter Corpus[2].

However, the following sentence has been published in the proceedings of LREC2004.

- For instance, WordNet has been used for <u>text summarization</u>.

[1] http://wordnet.princeton.edu/wordnet/publications/

[2] http://about.reuters.com/researchandstandards/corpus/

This shows that WordNet is useful for text summarization. Moreover, the following sentence has been published in the proceedings of CoNLL2003.

- In this paper, Long Short-Term Memory is applied to named entity recognition, using data from the Reuters Corpus.

This shows that Reuter Corpus is useful for named entity recognition. These information are not announced on official home pages.

As mentioned above, lists of usage information described in academic articles have ones that are not originally considered by developer. That is, if we could extract usage information for LRs from a large number of articles, we could efficiently find and use appropriate LRs.

3 Usefulness of Automatically Extracted Usage Information

The aim of our research is to enable more LRs suitable for user's purpose to be searched by automatically extracting usage information. In this section, we show that automatically extracted usage information contributes to the LR searches. In particular, we constructed usage information database by combining lists of usage information based on developer's perspective manually extracted from official home pages with ones automatically extracted from academic articles. Then, we compared the number of LRs retrieved using queries with and without automatically extracted usage information.

3.1 Construction of Usage Information Database

Figure 1 illustrates the flow for creating the usage information database. We registered lists of usage information based on developer's perspective and automatically extracted usage information in the database. Metadata registered in SHACHI (Tohyama et al., 2008a) was used as usage information based on developer's perspective. SHACHI is the metadata database of LRs, consisting of approximately 2,400 pieces of meta information. Usage information manually extracted from official home pages has been described in the 'type.purpose,' which is one of 45 metadata (Tohyama et al., 2008b) in SHACHI. While, we used usage information extracted from LREC2004 and LREC2006 by the extraction rules (Kozawa et al., 2008) as automatically extracted usage information. 728 sentences containing usage information for 413 LRs were extracted by using the extraction rules and registered in the database.

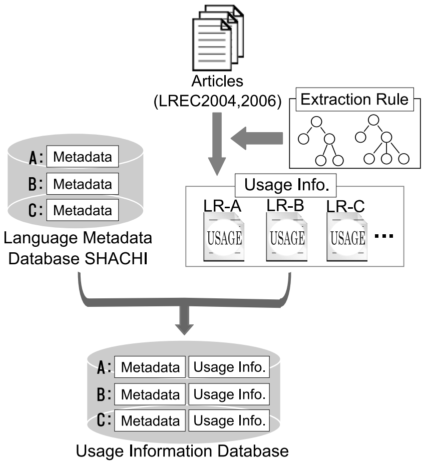

Figure 1: Flow for creating database of usage information

3.2 Verification Experiment by LR Searches

We performed experiments to know the usefulness of the automatically extracted usage information. We verified whether the number of LRs suitable for user's needs is increased by combining the automatically extracted usage information with the metadata database. In the experiments, we used keywords as queries and got a list of LRs whose the metadata 'type.purpose' or automatically extracted usage information contained the keywords.

3.2.1 Search Experiment Using Keywords in Research Topics

We carried out experiments searching for LRs on the database using keywords to learn whether the number of retrieved LRs increases by using the automatically extracted usage information. As queries for the LR search, we used 40 keywords in the ACL2008 Call for Papers since we assumed that researchers in the fields of computational linguistics searches for LRs suitable for their own purposes. We compared the number of LRs using retrieved both the 'type.purpose' and the extracted usage information as index with one using only the 'type.purpose.' However, when the automatically extracted usage information were used, we judged the validity between queries and LRs contained in search results. This is because search results have possibilities that they contained invalid LRs due to mistakes by the extraction rules.

The experimental results are shown in Table

Table 1: Results of verification experiment for automatically extracted usage information

	type.purpose		type.purpose & Usage Info.	
Keyword	Keyword	Keyword & Synonym	Keyword	Precision (%)
dialogue	8	10	10	100 (10/10)
information retrieval	52	57	65	100 (65/65)
information extraction	11	26	26	100 (26/26)
question answering	0	1	16	100 (16/16)
summarization	8	8	23	85.2 (23/27)
machine translation	55	55	56	100 (56/56)
speech recognition	211	230	280	96.6 (280/290)
speech synthesis	32	37	40	95.2 (40/42)
syntax	0	3	0	0 (0/34)
semantics	1	4	1	33.3 (1/3)
word sense disambiguation	0	2	15	100 (15/15)
discourse	10	10	17	51.5 (17/33)

1. The number of LRs retrieved using only the ‘type.purpose’ is shown in column 2 and the number of valid LRs retrieved using both the ‘type.purpose’ and the extracted usage information is shown in column 4. The retrieval precision using both the ‘type.purpose’ and the extracted usage information is shown in column 5. The number of retrieved LRs increased for seven keywords by adding the extracted usage information. This indicates that the automatically extracted usage information is useful for LR searches.

In this research, for the purpose of searching LRs broadly, we performed the expansion of indexes by using the extracted usage information. However, as another approach, the expansion of queries should be also considered. We showed that whether LRs that could not be retrieved by only expanding queries exists or not if we compared LRs retrieved by the expansion of indexes with ones retrieved by the expansion of queries. Therefore, we searched for LRs from the ‘type.purpose’ using synonyms of 40 keywords. Synonyms were generated by hand based on the analysis of ‘type.purpose.’ The synonyms used in the experiment were designed to fill surface or semantic gaps.

In comparison with the number of LRs retrieved using synonyms (column 3 & 4), the number using the automatically extracted usage information increased as for “information retrieval”, “question answering” and so on. These results show that LRs which could not be retrieved by only expanding queries exists.

Table 2: Results of subject experiment

Subject ID	1	2	3	4	5	6	7
retrieved LRs without usage information	5	2	3	2	5	2	4
retrieved LRs by adding usage information	2	1	0	1	1	2	2

3.2.2 Subject Experiment

In Section 3.2.1, we evaluated the number of LRs matched with queries. However, all retrieved LRs are not suitable for user’s needs. Therefore, we carried out subject experiments to confirm the usefulness of the automatically extracted usage information. The subjects searched LRs suitable for a research purpose. We checked whether subjects found LRs that can not search using only the ‘type.purpose,’ by adding the extracted usage information to indexes. In the experiment, seven subjects who engaged in the field of computational linguistics made a 10-minute LR search. Subjects 1 to 3 searched for LRs suitable for information retrieval, subjects 4 and 5 searched for LRs suitable for summarization, and subjects 6 and 7 searched for LRs suitable for question answering. The subjects did keyword searches and selected appropriate LRs from the search results by referring to metadata ‘description’ in order to judge whether they are suitable for their own purposes. The explanation of LRs is recorded in the ‘description.’

The experimental results are shown in Table 2. Consequently, by the searches with usage information, six among the seven subjects successfully found LRs that were overlooked by the searches without usage information. In addition, we conducted a questionnaire on LR search and found that six subjects answered that searching with usage information was more efficient than without. These results show that automatically extracted usage information has the potential to contribute to efficient LR searches.

3.3 Verification Experiment by Checking for Usage Information Database

When LRs were retrieved using the ‘type.purpose’ and the extracted usage information as indexes and 40 keywords and their synonyms as queries, the number of retrieved LRs was only 465, although approximately 2,400 LRs were registered in SHACHI. More LRs have to be able to be retrieved since the aim of our research is to develop availability of LRs by searching more LRs.

There are two reasons not to be able to retrieve:

(1) No query was matched with usage information

(2) usage information was not registered

We performed a sampling investigation using 100 LRs as sample to investigate why LRs were not retrieved. Out of 100 LRs, 35 LRs had been retrieved. 12 LRs were not retrieved for the reason (1) and other 53 LRs were not retrieved for the reason (2). In this section, we investigate the causation of (1) and how to deal with (2).

3.3.1 Investigation of Cause of Low Coverage

We investigate 12 LRs which were not retrieved although the 'type.purpose' or the automatically extracted usage information were registered in the database. There were 11 LRs that were not retrieved even though the 'type.purpose' was registered. The 'type.purpose' of these LRs contained medical and educational keywords. These keywords were different from the keywords in Table 1. While there was 1 LR that were not retrieved even though the extracted usage information was registered. The usage information for this LR, the "Oxford English Dictionary," contained "reducing the granularity of the WordNet sense inventory," which means word sense disambiguation. However, our thesaurus does not cover this usage. The reason why above 12 LRs were not retrieved was that keywords used in ACL2008 could not cover usages for LRs. However, these LRs could be retrieved by using other keywords.

3.3.2 Effect of Automatic Extraction of Usage Information

Registration of usage information is required for retrieving 53 LRs whose the 'type.purpose' and usage information were not registered in the database. We investigated possibilities that the number of retrieved LRs were increased by augmenting information resources for extracting usage information.

We applied the extraction rules to the proceedings of ICSLP2004 and ICSLP2006. ICSLP which is an international conference on spoken language processing, is a different field of conference from LREC. Five of 28 LRs were published in ICSLP2004 or ICSLP2006 and the usage information for four of them was extracted successfully. This indicates that we could extract usage information for LRs by applying the extraction rules to various resources.

4 Conclusion

In this paper, we described how automatically extracted lists of usage information for each LR contributed to efficient LR searches. By combining metadata databases of LRs with automatically extracted lists of usage information, we found that appropriate LRs could be searched more efficiently with usage information than without it.

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., 2008a).

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*.

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

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 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. 2008a. Construction of an infrastructure for providing users with suitable language resources. In *Proceedings of the 22nd International Conference on Computational Linguistics*, pages 119–122, http://shachi.org/.

Hitomi Tohyama, Shunsuke Kozawa, Kiyotaka Uchimoto, Shigeki Matsubara, and Hitoshi Isahara. 2008b. Shachi: A large scale metadata database of language resources. In *Proceedings of the 1st International Conference on Global Interoperability for Language Resources*, pages 205–212.