A Preliminary Analysis of the Use of Resources in Intelligent Information Access Research

Jiangping Chen

School of Library and Information Sciences, University of North Texas, P. O. Box 311068, Denton, TX 76203-1068 jpchen@unt.edu

Fei Li

School of Library and Information Sciences, University of North Texas, P. O. Box 311068, Denton, TX 76203-1068 fl0030@unt.edu

Cong Xuan

School of Library and Information Sciences, University of North Texas, P. O. Box 311068, Denton, TX 76203-1068 cx0050@unt.edu

This paper reports our exploratory analysis of the use of resources in three Intelligent Information Access (IIA) research areas: Automatic Classification, Question Answering, and Cross-Language Information Retrieval. Forty-three recently peer-reviewed papers from three annual conferences (SIGIR, ACL, and HLT) were selected and analyzed. The purpose of this analysis is twofold: 1) to explore methodological issues for large-scale content analysis of resources used in IIA research, and 2) to achieve a basic understanding of various ways that resources can be used in the three IIA subfields. The work reported in this paper is part of an effort to systematically explore the information needs for resources in Intelligent Information Access research.

Introduction

Intelligent Information Access (IIA) is a term that has been used (Berry, Dumais & Letsche, 1995; Maybury, 2005; Müller, 1999) but not clearly defined. In this paper, Intelligent Information Access (IIA) refers to technologies that makes use of human knowledge or human-like intelligence to provide effective and efficient access to large, distributed, heterogeneous and multilingual (and at this time mainly text-based) information resources and to satisfy users' information needs. In other words, any information access technologies

involving applying human knowledge to retrieve/understand/synthesize or extract information are considered as Intelligent Information Access. Particularly, IIA includes technologies on Automatic Classification and Clustering, Summarization, Information Extraction, Cross-Language Information Retrieval (CLIR), as well as Question Answering (QA).

Recent IIA research has a common characteristic: researchers typically make use of various knowledge resources and system tools in order to design, construct, and evaluate their systems (Chen, et al., 2004; Kishida, et al., 2004). Part of the reason for this situation is due to the fact that the tasks underlying these fields are normally complex and require sophisticated application of human knowledge. This has the unfortunate effect of forcing the research team to spend time and effort on developing appropriate knowledge resources themselves.

The advent of the Internet has enabled global collaboration and resource sharing. More and more researchers in the IIA field choose to make use of freely available resources on the Web to expedite system development or to facilitate system evaluation. Here, resource is broadly defined as any knowledge resources such as ontology, annotated corpus, test collections and various software systems such as information retrieval (IR) systems or search engines, machine translation systems, and various natural language processing (NLP) tools.

Today, as more and more scholars in computer science, library science, and information science become interested in IIA research or applying IIA technologies for educational purposes, there is an increasing need for resource sharing in IIA, both to establish common ground for evaluating and comparing approaches and to open the playing field to smaller research groups. However, there is a scarcity of analyses and reviews in the current status of the use of resources in IIA. As a result, there is no systematic collection and/or annotation of resources that have been used in the various IIA fields. Individual researchers have to conduct time-consuming literature review or information seeking to investigate whether there is any appropriate resources available, or they have to make the effort to create the resources they need for their research.

We are trying to change this situation by conducting a research project regarding the use of resources for IIA. The objectives of our project include: 1) to discover what kinds of resources have been used in IIA research and how they have been used; and 2) to investigate IIA researcher's information needs and the challenges they have found as it relates to the use of resources in their research.

The research is exploratory since no similar investigation has been carried out before. We decided to divide the whole project into three phases. Each phase has its own research

questions and objectives. In Phase One, a small set of IIA research papers are selected and analyzed. The purpose is to determine appropriate analytical approaches and to obtain a general understanding of the IIA resource reality. A coding scheme will be developed, evaluated, and revised so that it can be used for similar studies; In Phase Two, more research papers will be analyzed applying the coding scheme as created from Phase One. After the analysis in this phase, we expect to discover what kinds of resources have been used in IIA research and how they have been used. Here, a web-based database system will be designed to record the resources. In Phase Three, an online survey will be carried out to collect problems and information needs regarding the use of various resources from the IIA community. Questionnaires will also be sent to researchers who have published at least one research paper or report in IIA. Through an analysis of their responses, we expect to gain an in-depth understanding of the use of resources as they are used in the real world. It is expected that this will inform and guide the development and use of resources for IIA research and applications in the future.

This paper reports on the work conducted in Phase One. Previous academic research and practical efforts regarding the use of resources for IIA is reviewed in the next section. Following this are the methodological issues for data collection and analysis. Then, the results of analysis are presented. The paper concludes with a summary of findings and lessons learned from the analysis.

Literature Review

Among the five IIA subfields as specified by Maybury (2005) including Information Retrieval, Summarization, Information Extraction, Text Clustering, and Question Answering, some are well investigated both theoretically and practically. Significant research forums such as Text REtrieval Conference (TREC), Cross-Language Evaluation Forum (CLEF), and NII-NACSIS Test Collection for IR Systems Workshop (NTCIR) conduct large-scale evaluation on tasks such as Cross-Language Information Retrieval (CLIR), monolingual Question and Answering (QA), and Cross-Language Question Answering (CLQA). Besides, The Association for Computational Linguistics (ACL) and ACM Special Interest Group in Information Retrieval hold conferences each year involving researchers from areas of QA, CLIR, Automatic Classification, and so forth.

No systematic review or summarization of resources that have been used in IIA has been conducted. However, there does exist some practical efforts to collect web resources that can be used for research or education in IIA. Kraft (n.d.) has located various information retrieval systems and projects, and has put the links on his website. The Scottish electronic Staff Development Library (SeSDL) has made available some vocabularies and thesauri (SeSDL, 2000). Tools and services for automatic classification are presented at

http://searchtools.com/info/classifiers-tools.html , with a brief annotation for each resource (Search Tools Consulting, 2003). Amitay (n.d.) has built the Web IR & IE site, which is a collection of online resources for research in Information Retrieval and Information Extraction. Resources in various formats for information retrieval education are provided by the School of Information Sciences, University of Pittsburgh (n.d.). These web pages allow researchers to review each individual resource following the link for their research. However, there is no indication of whether these resources have been used and for what purposes.

Some researchers have looked into the effect of certain resources in a particular subfield of IIA. For example, Lin (2002) compared two approaches to using Web data for Question Answering: a federated approach and distributed approach. Lita et al. (2004) quantified the utility of several types of widely-used resources in QA, including Web search engines, gazetteers, encyclopedias and so forth. Kosovac (n.d.) identified functions of thesauri in aiding indexing and searching Internet and intranet information based on certain examples. The above efforts do help others to evaluate certain resources; however, none of them provides a thorough review of the availability and use of resources in the discussed IIA subfield.

Research Plan and Methodological Issues

Research Design and Research Questions

As introduced before, this paper reports on the first phase of our research project which aims to systematically investigate various resources and their use in Intelligent Information Access. The focus of the first phase is to find out an appropriate analytical approach for large-scale analysis to be conducted in the second phase.

Content analysis (Krippendorff, 1980) approach is our natural choice for the analysis carried out through the project because resources and their use situation can be coded and extracted from IIA research papers describing IIA experiments and systems through content analysis. A small-scale content analysis was performed on 43 selected research papers and posters (see table 1) from the three top annual conferences held in 2005 in the field of IIA. The research questions we would like to explore include:

> (1) What is an appropriate coding scheme for analyzing the use of resources in IIA? and, (2) What are the characteristics of the analysis?

Since there is no existing coding scheme that can be adapted for use in our analysis. It is essential that we develop an appropriate coding scheme to guide the analysis throughout the whole project in the first phase. We would like to develop a coding scheme that is simple but inclusive. It should contain all important categories so that important

information regarding the use of resources in an IIA research paper can be discovered. The coding scheme should also facilitate the analysis process so that trained coders can easily locate the information and make objective judgments on classification during coding.

Another purpose of the first phase is to identify the basic characteristics and challenges of the content analysis process. The results will help us to achieve a better understanding of the methodology itself and to apply it to appropriate research settings. The results will also help us to design appropriate survey questions in Phase Three.

The following subsections discuss sampling criteria, instrument development and data collection procedures.

Sampling Criteria

To achieve our purposes, we selected three subfields as our starting points: Automatic Classification and Clustering, Question Answering (QA) including monolingual QA and Cross-Language Question Answering (CLQA), and Cross-Language Information Retrieval (CLIR). The reasons of choosing these three subfields are: 1) they have been well investigated and attract many researchers from different fields; and 2) research in these subfields is more likely to build on various resources developed by others. Considering the fact that in the IIA field, most new studies and experiments are first submitted to conferences rather than journals and there are several renowned international conferences holding special topics as well as experimental tasks in IIA each year, we decided only to include conference publications in our investigation.

This study seeks to understand the real-word status of current IIA resources; therefore, we gave priority to research papers/posters that were most recently published. Specifically, the sample consists of 43 papers/posters, selected from the proceedings of three conferences including ACM SIGIR (Special Interest Group on Information Retrieval) 2005 Annual Conference, 2005 Annual Conference of the Association for Computational Linguistics (ACL), and 2005 Human Language Technology Conference (HLT), which are among the major international conferences in IIA. We extracted all the papers/posters that are pertinent to any one of the three IIA subfields in the three conference proceedings. Table 1 lists the number of papers/posters selected from each of the three conferences and the distribution of the papers on the three subfields.

Table 1. Distribution of Sample Papers for Phase One

Conferences	QA	CLIR	Automatic Classification	Total
2005 SIGIR	4	5	10	19

2005 ACL	6	2	5	13
2005 HLT	7	1	3	11
Total	17	8	18	43

Preliminary Coding Scheme and Sheets

Chen (2003) designed a coding scheme for the analysis of translation resources used in TREC-9 CLIR systems, which was limited to CLIR. Referencing her coding scheme, we inductively developed a preliminary coding scheme from our pervious research experience in IIA as well as analysis of a few sample papers. As mentioned before, one purpose of this phase is to establish an appropriate coding scheme that can be used not only for our next research phase but also by other researchers in the future. The preliminary coding scheme was our starting point for development. Figure 1 presents the preliminary coding scheme we used to code the 43 papers.

The preliminary coding scheme contains five top categories: resource type, resource acquisition, subfield of use, category of use, and specific purpose of use. In Category A, "resource type" is concerned with the variety of the types of resources in IIA. This section is divided into lexicon/knowledge base/ontology/corpus and software systems. Category A.2 "software system" is further divided according to function. Category B "resources acquisition" lists the different possible approaches of acquiring a resource. Some resources might be freely available on the web or through other channels, and some might be purchased by the researcher. Those resources that were neither open to everyone nor purchased by researchers (e.g., TREC test collection) were categorized into B.3 “other”. Category C identifies the subfield the paper is in and hence is composed of four subcategories: monolingual QA, CLQA, CLIR, and Automatic Classification or Clustering. Category D differentiates how the resource might be used - whether the whole resource (e.g., a machine translation tool) or only some part of it (e.g., WordNet) might be used. The last category E was not well-developed. We plan to ask the coders to identify and record the specific purposes of use of each resource in each paper. To make sure that the coding scheme is inclusive, we added subcategories such as “Unclear” and “Other” under certain categories so that unexpected situations can be coded. We realized that names of some categories were not concise or accurate and need further examination, but we used them to start the analysis.

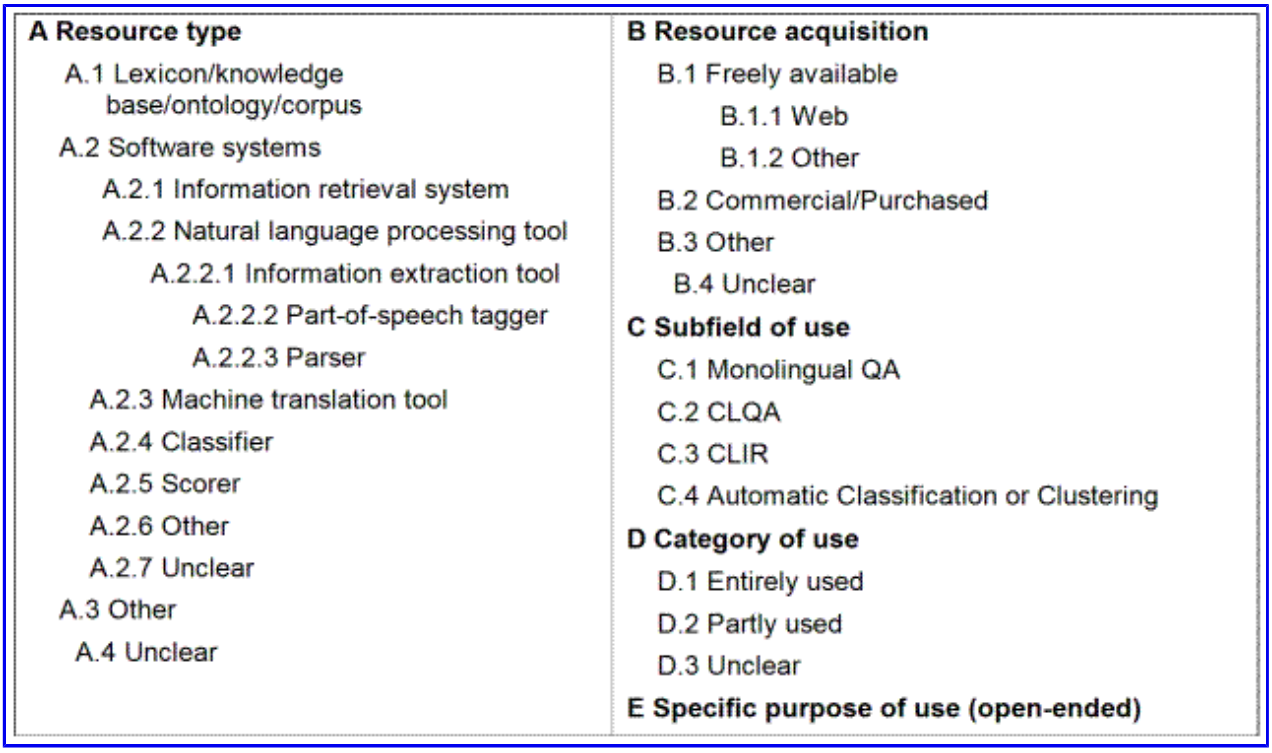

Figure 1. Preliminary coding scheme

Two coding sheets were designed according to the coding scheme. The first sheet records the use of each resource in each paper (i.e., Category C, D and E).The second one focuses on each resource and particularly records the type (i.e., Category A) and acquisition method (i.e., Category B) for each one.

Coding Criteria and Procedures

Resources that were developed by the original researchers themselves and used only for their own studies, were excluded from the analysis due to the difficulty of coding them and our feeling that it is more meaningful and valuable to the IIA community to focus on resources that have been used by other researchers for this study.

Prior to coding, the initial coding scheme was tested on several papers in the sample to determine the categories. Then, two coders (i.e., the second and the third authors of this paper) received instruction for the coding scheme, coding sheets, and judging criteria. Afterwards, they coded the 43 papers independently. Each resource was coded in Category A, B, C and D based on the paper. The coders used the web to originally clarify unclear categories. A resource was labeled “unclear” if its categories could be not verified by the paper or from the web.

The two coders were getting familiar with IIA literature through the coding process. Their

coding process was quite slow in the beginning because they had to understand many concepts associated with the use of resources in the papers. But once they became familiar with the subjects, the coding went more smoothly. Based on Krippendorff's alpha, a reliability coefficient developed to measure the agreement between coders (Krippendorff, n.d.), the intercoder reliability for Category A is .963 and for Category B is 1.000. It shows that there is little disagreement between the two coders. Finally, the three authors met together and discussed all the disagreements in the coding process, which then determined the final codes. Next, we present the coding results and our analysis.

Results and Discussion

Our preliminary study provides evidence that IIA research makes use of various resources including web resources. A list of 108 resources was identified. 51 out of the 108 (47.2%) resources are various software systems such as Information Retrieval systems, Natural Language Processing systems, and machine translation tools. 55 of the 108 resources (50.9%) belong to lexicon/ontology/ knowledge base/corpus. Over half of the resources (58/108) are clearly identified as Web resources. No resources are coded in C.2 CLQA, which shows that no resources were used in CLQA research in the sample.

Many sample papers do not clarify how the resources were acquired. The coders conducted searching on the web and have located some of them. But still, 41 out of 108 resources are unclear concerning the acquisition method. This issue could be addressed through the survey in Phase Three. A summary of the coding results is presented in Table 2.

Several drawbacks of the coding scheme were identified from the results. First and foremost, the large number of resources coded into A.1 lexicon/knowledge base/ontology/corpus implies the need to break it down into several subcategories. Second, since only one experiment used "scorers" and referred them in the paper without clear definition, it might be inappropriate to consider it as a subcategory. Third, through the analysis, we found morphological tools as an important sub-genre software and therefore, should be used as a subcategory of A.2.2 natural language processing tools. Forth, the current coding scheme is restricted to three subfields of IIA and thus is not applicable to the other subfields such as Summarization and Information Extraction.

Table 2. A Summary of Coding Results

Category	Frequency	Category	Frequency
A Resource type	**108**	**B Resource acquisition**	**108**
A.1 Lexicon/knowledge base/ontology/corpus	55	B.1 Freely available	59

A.2 Software system	51	B.1.1 Web	59
A.2.1 Information retrieval system	6	B.1.2 Other	0
A.2.2 Natural language processing tool	14	B.2 Purchased	3
A.2.2.1 Information extraction tool	2	B.3 Other	5
A.2.2.2 POS tagger	5	B.4 Unclear	41
A.2.2.3 Parser	7	**C Subfield of use**	**136**
A.2.3 Machine translation tool	3	C.1 Monolingual QA	49
A.2.4 Classifier	7	C.2 CLQA	0
A.2.5 Scorer	2	C.3 CLIR	25
A.2.6 Other	18	C.4 Automatic classification & clustering	62
A.2.7 Unclear	1	**D Category of use**	**136**
A.3 Other	1	D.1 entirely used	107
A.4 Unclear	1	D.2 partly used	19
		D.3 unclear	10
		E Specific purpose of Use	/

As seen below, we report some characteristics of the use of resources identified through the analysis and come up with a refined coding scheme for analyzing the use of resources in the IIA field.

Sample Distribution of Resource Use

The coding results show that most research utilizes resources developed by others. Among the 43 sample papers, only 5 (11%) of them did not use any resources that were developed by others. Most of the research used certain number of resources. Table 3 is the frequency distribution of papers on the use of resources. The majority papers (33 out of 43) mentioned the use of 2-6 resources.

Resource Types and Frequently Used Resources

A further examination of the coding scheme on Category A, resource type, shows that approximately half of the resources identified fall into A.1. lexicon/knowledge base/ontology/corpus. This indicates that this category should be further divided into several subcategories to provide a more refined categorization. Due to the limited sample

size, the duplication rate of the use of the same resource in different papers is low. Most of the resources (90/108, i.e., 83.3%) were only mentioned in a single paper.

Table 3. Frequency Distribution of Sample Papers

Number of resources used	Number of papers
None	5
1	3
2	10
3	8
4	8
5	3
6	4
7	0
8	1
9	0
10	1
total	43

According to the results, the most frequently used resources include TREC data collection, WordNet, SVM light, Reuter corpus, and ROUGE. Researchers used TREC data collection to test systems and construct corpora. WordNet has a broad usage, including semantic taxonomy construction, tagging, question ranking, and feature identification. SVM light was used for system comparison, system training, and word sense disambiguation. Reuter's corpus was primarily used to test systems. The usage of ROUGE includes system comparison and automatic evaluation.

Distribution of Resource Use in the Three Subfields

Table 4. Distribution of Resource Use among the Three Fields

Resource type	Frequency of use for QA	Frequency of use for CLIR	Frequency of use for Automatic Classification	Total frequency of use
Lexicon/knowledge base/ontology/corpus	23	14	35	72
Information retrieval system	7	2	0	9

Natural Language Processing tool	5	5	5	15
Machine translation tool	1	2	1	4
Classifier	3	0	8	11
Other	9	3	13	25
Total	48	26	62	136

It is interesting to observe the distribution of resource use among the three fields. Table 4 presents the results. It shows that although various IIA subfields use different information resources and tools, QA research typically uses all kinds of resources from ontology to classifier. However, CLIR research seldom uses classifiers, and few Automatic Classification research uses IR systems.

Web Resources

There are totally 59 freely accessible Web resources used in the sample papers. Nearly half of them are knowledge sources (including lexicon, knowledge base, ontology, and corpus). Web resources within each resource type are listed in Table 5.

Table 5. Web Resources of Each Resource Type

Resource Type	Number of resources	Resources
Knowledge source	27	WordNet, Reuters corpus, AQUAINT corpus, 20 Newsgroups dataset, LocusLink database, MEDLINE database, WebKB dataset, ACE conference corpus, ACM digital library documents, Encyclopedia.com definitions, English FrameNet, Europarl, Internet Movie Review Database archive, LDC dictionaries, Medical Subject Heading (MeSH), Metathesaurus of the Unified Medical Language System (UMLS), MPQA corpus, Polarity, Prague Czech-English Dependency Treebank, PubMed stopword list, SALSA database, SemCor, SENSEVAL-3 English lexical samples, SMART stoplist, UCI machine learning dataset, UIUC data set, Usenet newsgroups
IR system	5	LUCENE, Lemur toolkit, Google search engine, Altavista search engine, SMART system
IE system	1	CRF package
POS tagger	4	FreeLing, MontyTagger, TnT tagger, Tree tagger
Parser	3	Minipar, CASS partial parser, Stanford Lexicalized Parser

Machine translation tool	3	GIZA, GIZA++, Google translator
Classifier	6	LIBSVM package, SVM ^{light}, ACM Computing Classification System, BNT package, SVMTorch, TextCat
Other and Unclear software system	8	ROUGE, HP-Filter, Minorthird, MORPHOSAURUS text processing engine, PubMed, SRILM toolkit, Entrez Utilities, FSA
Other and Unclear resource type	2	HighWire DTD, Open Directory project data

Refined coding scheme

According to the deficiencies discovered through the data analysis process, we further refined our coding scheme, as presented in Figure 2.

Compared to the preliminary scheme, the revision includes the following: (1) break Category A.1 down into monolingual lexicon/ontology, bilingual or multilingual lexicon/ontology, annotated corpus, un-annotated corpus, and test collection and change the name for Category A.1 into “Knowledge source”; (2) add a category “B.3 self-developed” based on the consideration that self-developed resources still have potential for future IIA research; (3) add “morphological processor” and “other” under A.2.2 natural language processing tools; (4) remove the category “scorer” under A.2 software system; (5) add the other IIA subfields (i.e., Summarization and Information Extraction) to the scheme; (6) combine the previous Category C, D and E into a single category called “Usage”, including three subcategories: Subfield, Proportion of use, and Specific purpose of use; and (7) further develop the “Specific purpose of use” category based on what has been discovered in the analysis. Due to the small sample size in Phase One, these items might not be inclusive and need further development through the next phase. Therefore categories “3.16 Other” and “3.17 unclear” are added.

A Resource type

- A.1 Knowledge source
  - A.1.1 Monolingual lexicon/ontology
  - A.1.2 Bilingual or multilingual lexicon/ontology
  - A.1.3 Annotated corpus
  - A.1.4 Unannotated corpus
  - A.1.5 Test collection
  - A.1.6 Other
  - A.1.7 Unclear
- A.2 Software system
  - A.2.1 Information retrieval system
  - A.2.2 Natural language processing tool
    - A.2.2.1 Information extraction tool
    - A.2.2.2 Part-of-speech tagger
    - A.2.2.3 Parser
    - A.2.2.4 Morphological processor
    - A.2.2.5 Other
  - A.2.3 Machine translation tool
  - A.2.4 Classifier
  - A.2.5 Other
  - A.2.6 Unclear
- A.3 Other
- A.4 Unclear

B Resource acquisition

- B.1 Freely available
  - B.1.1 Web
  - B.1.2 Other
- B.2 Purchased
- B.3 Self-Developed
- B.4 Other
- B.5 Unclear

C Usage

- C.1 Subfield
  - C.1.1 Monolingual QA
  - C.1.2 CLQA
  - C.1.3 CLIR
  - C.1.4 Automatic Classification or Clustering
  - C.1.5 Summarization
  - C.1.6 Information Extraction
  - C.1.7 Other
- C.2 Proportion of use
  - C.2.1 Entirely used
  - C.2.2 Partly used
  - C.2.3 Unclear
- C.3 Specific purpose of use
  - C.3.1 Stop word removal
  - C.3.2 Indexing
  - C.3.3 Query expansion
  - C.3.4 Document retrieval
  - C.3.5 Disambiguation
  - C.3.6 Morphological analysis
  - C.3.7 Parsing
  - C.3.8 Tagging
  - C.3.9 Term translation
  - C.3.10 Named entity identification
  - C.3.11 Corpus construction
  - C.3.12 System or corpus training
  - C.3.13 Ranking
  - C.3.14 Answer extraction
  - C.3.15 Classification
  - C.3.16 Other
  - C.3.17 Unclear

Figure 2. Refined coding scheme

Conclusion

Our preliminary analysis in the use of resources in the IIA field shows that most IIA research involves resources that are developed by other researchers and the majority of resources are freely available on the web. A number of types of resources were utilized, such as test collections, annotated corpus, and many different kinds of software systems. Among all the resources identified in the sample papers, half of them are knowledge sources (including lexicon, ontology, knowledge base, and corpus) and nearly another half belonging to

software systems. Different subfields, given their different focuses and purposes of research, may have emphasis on different types of resources.

The results from the analysis in Phase One helped us discover the drawbacks and limitations of the sampling method and the preliminary coding scheme. The revised coding scheme is more complete and can be used in large scale analysis in all subfields. The scheme breaks Category A.1 into several subcategories providing a more detailed picture in the use of different sorts of knowledge sources in the IIA field. In addition, the refined scheme tentatively develops the "specific purpose of use", based on the data gathered during this phase. The categorization of this part might not be inclusive and needs further testing and modification on a larger group of samples.

We believe that the content analysis approach is appropriate in discovering the big picture of the use of resources in IIA. However, we have noticed that the analysis has limitations towards understanding the degree of satisfaction of IIA researchers on the resource they choose to use because the information is normally not provided in research papers. The survey research to be conducted in Phase Three should provide more information on this issue. Altogether, our study will help IIA community and other fields to better develop, share, and use resources for IIA research and system development.

References

Amitay, E. (n.d.) Web IR & IE Retrieved June, 5, 2006, from http://www.webir.org/

Berry, M. W., Dumais, S. T., & Letsche, T. A. (1995) Computational methods for intelligent information access *Proceedings of Supercomputing'95, San Diego, CA, December 1995* Retrieved February 12, 2006, from http://citeseer.ist.psu.edu/berry95computational.html

Chen, J. (2003) *The construction, use, and evaluation of a lexical knowledge base for English-Chinese cross language information retrieval* Ph.D. Dissertation. Syracuse University

Chen, J., Ge, H., Wu, Y., & Jiang, J. (2004) UNT at TREC 2004: Question Answering Combining Multiple Evidences *Proceedings of TREC 2004* Retrieved October 10, 2005, from http://trec.nist.gov/pubs/trec13/papers/unorthtexas.qa.pdf

Kishida, K., Chen, K., Lee, S., Kuriyama, K., Kando, N., Chen, H.-H., Myaeng, S. et al. (2004) Overview of CLIR Task at the Fourth NTCIR Workshop *Proceedings of NTCIR-4* Retrieved February 12, 2006, from http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings4/CLIR/NTCIR4-CLIR-TaskOverview.pdf

Kraft, D. (n.d.) *Information Retrieval Systems* Retrieved June 6, 2006, from http://bit.csc.lsu.edu/~kraft/retrieval.html

Krippendorff, K. (1980) *Content Analysis and Its Methodology* Beverly Hills, CA: SAGE Publications

Krippendorff, K. (n.d.) Computing Krippendorff's Alpha-Reliability Retrieved May 30, 2006, from http://www.asc.upenn.edu/usr/krippendorff/webreliability2.pdf

Kosovac, B. (n.d.) *Internet/Intranet and thesauri* Retrieved January 9, 2006, from http://irc.nrc-cnrc.gc.ca/thesaurus/roofing/report_b.html

Lin, J. (2002) The Web as a resource for Question Answering: perspectives and challenges *Proceedings of the third International Conference on Language Resource and Evaluation (LREC 2002), Canary Islands, Spain* Retrieved January 10, 2006, from http://www.umiacs.umd.edu/~jimmylin/publications/Lin-LREC02.pdf

Lita, L.V., Hunt, W.A., & Nyberg, E. (2004) *Resource analysis for question answering* Retrieved January 9, 2006 from http://acl.ldc.upenn.edu/acl2004/postersdemos/pdf/lita.pdf

Maybury, T.M. (2005) Intelligent information access: theory and practice *Proceedings of 2005 International Conference on Intelligence Analysis* Retrieved November 20, 2005, from https://analysis.mitre.org/proceedings/Final_Papers_Files/272_Camera_Ready_Paper.pdf

Müller, M. E. (1999) Intelligent information access in the Web: ML based user modeling for high precision meta-search *Proceedings of the Workshop on Machine Learning for Intelligent Information Access, The ECCAI Advanced Course on Artificial Intelligence (ACAI-99), Chania, Greece, 1999* Retrieved February 12, 2006, from http://citeseer.ist.psu.edu/article/muller99intelligent.html

School of Information Sciences, University of Pittsburgh. (n.d.) *Information Retrieval Education Resources* Retrieved June 5, 2006, from http://ir.exp.sis.pitt.edu/res2/resources.php

Search Tools Consulting. (2003) *Tools for Taxonomies, Browsable Directories, and Classifying Documents into Categories* Retrieved June 6, 2006, from http://www.searchtools.com/info/classifiers-tools.html

SeSDL. (2000) *Taxonomy, Classification and Metadata Resource* Retrieved June 6, 2006, from http://www.sesdl.scotcit.ac.uk/taxonomy_links.html