IREX: IR and IE Evaluation project in Japanese

Satoshi Sekine[*]**, Hitoshi Isahara**[†]

[*]Computer Science Department
New York University
715 Broadway, 7th floor
New York, NY 10003 USA
sekine@cs.nyu.edu

[†]KARC, CRL
588-2 Iwaoka, Iwaoka-chou, Nishi-ku, Kobe
HYOUGO 651-2401, JAPAN
isahara@crl.go.jp

Abstract

We will report on the IREX (Information Retrieval and Extraction Exercise) project. It is an evaluation-based project for Information Retrieval and Information Extraction in Japanese. The project started in May 1998 and concluded in September 1999 with the IREX workshop held in Tokyo with more than 150 attendance (IREX_ Commettee, 1999). There is a homepage of the project at (IREX, Homepage) and anyone can download almost all the data and the tools produced by the project for free.

1. Background

Needless to say, the need for IR and IE technologies is getting larger because of the improvements in computer technology and the appearance of the Internet. In particular, it is really hard to find useful information from large quantities of electronic documents, such as newspapers and homepages. Because of this situation, research on Information Retrieval and Information Extraction is being actively conducted all over the world. Many researchers in the field feel that the the evaluation-based projects in the USA, MUC(MUC, Homepage) and TREC(TREC, Homepage), have played a very important role in the field. In Japan, however, there has been good research, but we have had some difficulties comparing systems based on the same platform, since our research is conducted at many different universities, companies, and laboratories using different data and evaluation measure. Our goal is to have a common platform in order to evaluate systems with the same standard. We believe such projects are useful not only for comparing system performance but also to address the following issues:

- To share and exchange problems among researchers.
- To accumulate large quantities of data.
- To let other people know the importance and the quality of Information Retrieval and Information Extraction techniques.
- To attract young researchers into the field.
- To start a long term and larger-size project of this kind.

The IREX project called for participants who share such goals. The project has been conducted under an open environment mainly based on mailing-list discussions and by volunteers.

2. Tasks

There were two tasks in IREX. Anyone can participate in one or both tasks.

- Information Retrieval task (IR)
  IR is the task of retrieving documents relevant to a given topic from a database of newspaper articles. Each topic is expressed by a description using a few noun phrases and a narrative using a few sentences. The dataset to be retrieved consisted of two years of the Mainichi Newspaper (1994 and 1995), which is available to anyone at a reasonable price. The total number of articles in the dataset was 211,853 articles. At the formal run, there were 30 topics and for each topic, participants were requested to submit up to 300 articles in the order of confidence.
- Named Entity task (NE)
  NE is the task to extract Named Entities, such as names of organizations, persons, locations, and artifacts, time and numeric expressions, money and percentage expressions. At the evaluation, participants were asked to identify NE expressions with SGML tags as correct as possible. There were two types of runs at the formal run: one was on a restricted domain and the other was on an unrestricted domain.

We set the two tasks, IR and NE, in order to evaluate the basic techniques. For example, in IR and IE, it is also important to design a good user interface and to extract the user's intention. However, the tasks of this project should be closely related to these techniques and the data accumulated in the project must be useful for improving such techniques.

We conducted a survey of the participant's systems. One can see the system descriptions in the survey and determine the corresponding system performance using the system's ID. Doing this, one might be able to find out what kind of techniques lead to better performance. There were about 100 items in each IR and NE survey.

3. Participants

There were 15 and 14 participants from Japan and US for IR and IE task, respectively, and the total number of collaborators, including those performing IR judgments, and creating the NE definition and the answers, were 45. It was one of the first large evaluation projects on IR and IE in Japan. (All participants are shown in Table 1)

Participants for IR

Gifu Univ., Kyoto Univ., Tsukuba Univ., Tsuda-juku Univ., Tokushima Univ., Univ. of Library and Information Science, Toyohashi Univ. of Technology, Nara Inst. of Science and Technology, Communication Research Lab., National Center for Science Information Systems, Agriculture, Forestry and Fisheries Research Information Center, AdIn Research Inc., AT&T, NEC, Justsystem, Sharp, Toshiba, Matsushita Electric Industrial Co.

Participants for NE

Ibaraki Univ., Kyoto Univ., The Univ. of Tokyo, Toyohashi Univ. of Technology, Yokohama National Univ., New York University, Inst. of Behavioral Sciences, Communication Research Lab., NTT-A, NTT-B, NEC, Fujitsu Lab.-A, Fujitsu Lab.-B, Matsushita Electric Industrial Co., Teragram

Participants for IR judgment

Ibaraki Univ., Kyushu Inst. of Technology, The Univ. of Tokyo, Tokyo Inst. of Technology, Japan Advanced Inst. of Science and Technologies, Yokohama National Univ., New York University, Inst. of Behavioral Sciences, The National Language Research Inst., NTT data, SONY CSL, Oki Electric Industrial Co., Hitachi Ltd, IBM Japan, RICOH

Other participants

The Mainichi Newspapers, NIST, Kyushu Univ., Telecommunications Advancement Org. of Japan, Electrotechnical Lab., Advanced Telecommunications Research Inst., Mitsubishi Electric Co.

Table 1: IREX Participants

4. Schedule

Table 2 shows the project schedule. We mainly used e-mail (mailing-lists) for the discussions. This is partially because one of the co-chairman was physically apart from most of the participants, but we would like to mention that a project of this size can be successfully conducted without meeting very often.

May 29, ’98	The first meeting (Tokyo)
June 30, ’98	Distribute draft of definitions
July 31, ’98	Initial call for participation
Aug.13, ’98	Unofficial meeting at COLING
Sept.16, ’98	The second meeting (Tokyo)
Oct.16, ’98	Close discussion of NE definition
==Dry Run==	
Nov.9, ’98	Start IR dry run
Nov.16, ’98	End IR dry run
Nov.17, ’98	Start NE dry run
Nov.20, ’98	End NE dry run
Nov.30, ’98	The third meeting (Tokyo)
Feb.14, ’99	Distribute CRL NE data
Mar.15,’99	Final call for participation
==Formal run==	
Mar.13,’99	Distribute restricted domain of NE
April 5, ’99	Start IR formal run
April 12, ’99	End IR formal run
April 13, ’99	Freeze NE system development
May 13, ’99	Start NE formal run
May 17, ’99	End NE formal run
Sept.1, ’99	NTCIR/IREX joint workshop
Sept.2-3, ’99	IREX workshop

Table 2: Schedule

5. IR

IR is the task of retrieving documents relevant to a given topic from a set of newspaper articles. We used Mainichi newspaper articles from ’94 and ’95 on a CDROM. There were bugs in the data, i.e. there were duplicated article ID’s in two day’s articles (August 23 and 24, 1995), so all the articles of these days were excluded from the evaluation. The total number of articles was 211,853, as shown in Table 3.

At the formal run, each participant can submit two systems. For each topic, systems are asked to submit up to 300 articles in the order of confidence. A topic consists of the following two pieces of information. Systems can use any part or all of this information.

Description: Simple expression of the topic. Normally a compound noun with modifier. It consists of at most three content words.

Data	Number of articles
’94	101,058
’95	111,497
Aug.23, ’95	-366
Aug.24, ’95	-336
Total	211,853

Table 3: IR: Number of articles

Narrative: Explanation of the topic so a human can unambiguously judge as much as possible. It consists of two or three sentences, and if necessary, it can have dictionary-like explanations, synonyms and examples.

The following is an example of a topic.

```
<TOPIC>
<TOPIC-ID>1001</TOPIC-ID>
<DESCRIPTION>Corporate merging
</DESCRIPTION>
<NARRATIVE>The article describes a
corporate merging and in the article,
the name of the companies have to be
identifiable.  Information including
the field and the purpose of the
merging have to be identifiable.
Corporate merging includes corporate
acquisition, corporate unifications and
corporate buying.</NARRATIVE>
</TOPIC>
```

There were 6 topics in the dry run and 30 topics in the formal run. Judgment was done from all the articles submitted from the participants (pooling). At first, two student judges made judgments and basically only the articles which did not get the same judgment were judged by the final judge. Final judges are volunteers from the groups which did not participate in the IR formal run. There are three judgments, A, B and C. These are defined as follows:

A : The subject of the article matches the topic.

B : The subject of the article does not match, but a part of the article matches the topic. There are some relationships between the articles and the topic.

C : No relationship between the article and the topic.

The number of articles to be judged and the numbers of A and B judgments by the final judge for each topic are shown in Table 4. The number of participants in the dry run was 7 groups and 10 systems. The number of participants in the formal run was 15 groups and 22 systems.

The evaluation of the system performance was conducted using the trec_eval program, which was also used in the TREC project. This program can be downloaded from Cornell University by ftp (TREC_EVAL, FTP site).

The results were not open in the dry run, but were anonymously open (using randomly assigned system ID’s) in the formal run. Table 5 shows the highest, median and lowest scores of R-Precision at the dry run. R-Precision is only one of several IR evaluation measurements, but since it is a single value, R-Precision is used in this paper. R-Precision measures precision (or recall, they’re the same) after R docs have been retrieved, where R is the total number of relevant docs for a query. Thus if a query has 20 relevant docs, then precision is measured after 20 docs, while if it has 200 relevant docs, precision is measured after 200 docs. In the table, ”Answer=A” means that only the articles judged as ”A” are considered answers (relevant articles) and ”Answer=A&B” for ”A” and ”B” are considered answers.

System	Answer=A	Answer=A&B
Best	0.3913	0.5504
Median (5th)	0.2513	0.3675
Worst	0.1205	0.1857

Table 5: IR dry run result

Table 6 shows the evaluation result of the IR formal run. As the participants of the dry run and the formal run are not completely overlapped, it is difficult to compare. However, when ”Answer=A”, the results of the formal run are generally better than that of the dry run. When ”Answer=A&B”, the best score in the dry run is better than that of the formal run.

6. NE

NE is the task of extracting Named Entities, such as organization names, person names, location names, time expressions, or numeric expressions. It is one of the basic techniques in IR and IE. The definition of NE’s is described in an 18-page document (which is available through the IREX homepage). There are 8 kinds of NE’s shown in Table 7. At the exercise, participants were asked to tag NE expressions with the corresponding SGML tags as accurately as possible. We also introduced a tag “OPTIONAL” to help in cases where even a human could not tag unambiguously. If a system tags an expression within the OPTIONAL tag, it is just ignored for the scoring. However, if a system tags an expression across the beginning or ending tags, then it is considered an overgenerated tag. The process of making the definition was not easy, which was partially reported in (Satoshi Sekine, 1999). There were long and active discussions on this subject at the meeting and in the IREX mailing-list.

There were three kinds of NE exercises, the dry run, a restricted domain formal run, and an unrestricted domain formal run. Also we supplied three kinds of training data: the dry run training data, the CRL_NE data and the formal run domain restricted training data. Table 8 shows the size of each data set. Note that CRL_NE data belongs to the Communication Research Laboratory (CRL), but it is included in the table, because the data was created by IREX participants, using the definition of IREX-NE, and distributed through IREX.

8 groups and 11 systems participated in the dry run exercise. The articles were selected from 1994 Mainichi newspaper articles. The domain of the articles was chosen to be

TOPIC ID	A	B	# of art. judged
Dry Run			
1001	80	145	931
1002	89	61	1096
1003	42	407	1316
1004	108	66	1480
1005	50	41	1099
1006	66	77	1356
Formal run			
1007	175	300	2246
1008	29	73	2565
1009	99	125	1588
1010	14	29	2222
1011	88	158	2130
1012	25	42	1535
1013	199	260	1308
1014	141	260	1473
1015	132	176	1505
1016	43	45	2446
1017	20	81	2248
1018	55	101	2086
1019	42	45	1859
1020	94	173	1291
1021	58	68	2030
1022	19	31	2015
1023	33	68	2853
1024	60	74	2934
1025	67	138	2047
1026	72	165	1914
1027	65	165	2513
1028	100	115	2806
1029	23	62	1878
1030	92	121	2053
1031	109	178	2134
1032	44	78	2268
1033	9	49	2989
1034	60	131	1911
1035	53	88	2008
1036	32	88	2299

Table 4: Number of articles judged/A/B

System ID	Ans.=A	Ans.=A&B
1103a	0.4512	0.4882
1103b	0.4667	0.5192
1106	0.2352	0.2110
1110	0.3335	0.4276
1112	0.2788	0.3340
1120	0.2707	0.3345
1122a	0.3803	0.4681
1122b	0.4032	0.4735
1126	0.0954	0.0883
1128a	0.3388	0.3897
1128b	0.3917	0.4156
1132	0.0604	0.0792
1133a	0.2382	0.2282
1133b	0.2460	0.2248
1135a	0.4929	0.5102
1135b	0.4829	0.4868
1142	0.4456	0.4929
1144a	0.4656	0.5499
1144b	0.4592	0.5434
1145a	0.3350	0.3419
1145b	0.2544	0.2927
1146	0.2225	0.2744

Table 6: IR Formal run result

NE	Example
ORGANIZATION	The Diet, IREX Committee
PERSON	(Mr.)Obuchi, Wakanohana
LOCATION	Japan, Tokyo, Mt.Fuji,
ARTIFACT	Pentium Processor, Nobel Prize
DATE	September 2, 1999; Yesterday
TIME	11 PM, midnight
MONEY	100 yen, $12,345
PERCENT	10%, a half

Table 7: NE Classes

Data	# of articles
Dry Run training	46
Dry Run	36
CRL_NE data	1174
Formal run (restricted) training	23
Formal run (restricted)	20
Formal run (unrestricted)	71

Table 8: Data size

balanced, but we excluded articles with no sentences (for example, name listings of some sort). The evaluation results of the dry run were not distributed. Only the score of the best, median (6th out of 11 participants) and the worst results are reported, which is shown in Table 9.

In the formal run, in order to study system portability and the effect of domains on NE performance, we had two kinds of exercises: restricted domain and unrestricted domain. In the unrestricted domain exercise (general), we selected articles regardless of domain. We excluded articles with no sentences, as we did on the dry run. In order to ensure the fairness of the exercise, we used newspaper articles which no one had ever seen. We set the date to freeze the system development (April 13). The date for the evaluation

System	F-measure
Best	68.23
Median (6th)	58.39
Worst	17.41

Table 9: Dry Run result

was set one month after that date (May 13 to 17) so that we could select test articles from the period between those dates. We thank the Mainichi Newspaper Corporation for providing this data for us free of charge.

We distributed the domain of the domain restricted exercise about one month before the system freeze date. It was an ”arrest” domain defined as the following, (it was called ”arrest” as opposed to ”general”).

```
The articles are related to an event
"arrest".  The event is defined as the
arrest of a suspect or suspects by
police, National Police, State police
or other police forces including the
ones of foreign countries.  It includes
articles mentioning an arrest event in
the past.  It excludes articles which
have only information about requesting
an arrest warrant, an accusation or
sending the papers pertaining to a case
to an Attorney’s Office.
```

In the formal run, 14 groups and 15 systems participated in the exercise. The evaluation results are made public anonymously using system ID’s. Table 10 shows the evaluation results (F-measure) of the formal run. The score of the formal run is generally better than that of the dry run. Comparing the score of the general domain and arrest domain, 5 systems got a better score on the general domain and 10 systems got a better score on the arrests. We observed a big improvements from the systems which took the domain shift into account.

We found that there are three kinds of systems. One was pattern based systems, in which patterns are written by human. This can be a very laborious job, but the best performing system came from this category. Second was also pattern based systems, but the patterns are extracted from a tagged corpus by some automatic means. The last type was fully automatic systems which do not use explicit patterns. It is interesting to see that the top three systems came from the three categories; we can’t simply conclude which type of systems outperform others. The results and analyses are reported in detail in (Satoshi Sekine, 2000).

System ID	general	arrest
1201	57.69	54.17
1205	80.05	78.08
1213	66.60	59.87
1214	70.34	80.37
1215	66.74	74.56
1223	72.18	74.90
1224	75.30	77.61
1227	77.37	85.02
1229	57.63	64.81
1231	74.82	81.94
1234	71.96	72.77
1240	60.96	58.46
1247	83.86	87.43
1250a	69.82	70.12
1250b	57.76	55.24

Table 10: NE Formal run result

7. Future Project

There was another evaluation based project on IR in Japanese at the same time, which is called NTCIR (NTCIR, Homepage). Although these two projects run independently as the organization bodies were different, we had close relationship, for example, we had a joint workshop. Following the great success of the IREX and NTCIR and the encouragement by many participants, we decided to carry on the evaluation-based project. In the future, IREX and NTCIR run within the unified framework. We are currently preparing for IR, IE, and also for summarization task. The progressing discussion in the summarization project can be seen at their homepage (TSC, Homepage). The tentative schedule of the task is as follows: the first call-for-participation is out in April 2000, and the final workshop is scheduled for March 2001.

8. Summary

In this paper, we described the IREX (Information Retrieval and Extraction Exercise) project, which is an evaluation-based Information Retrieval and Information Extraction project. Since we had a lot of participants, we believe this project made an important impact on the field. We hope that some new directions will arise and many advances will be seen in the field based on the experiments and the discussions throughout the project. Also, we hope the data and the tools created by the project will be utilized by many people and will be useful to make a lot of improvements in the field.

Finally, we would like to mention that the success of the project owes to all of the participants of the project. We appreciate their participation and corporation.

9. References

IREX, Homepage. *http://cs.nyu.edu/cs/projects/proteus/irex*.

IREX_Commettee, 1999. *Proceedings of the IREX Workshop*.

MUC, Homepage. *http://www.muc.saic.com/*.

NTCIR, Homepage. *http://www.rd.nacsis.ac.jp/ñtcadm*.

Satoshi Sekine, Yoshio Eriguchi, 1999. Difficulties of ne definition–from the experiences in irex (in japanese). In *Proceedings of The Fifth Annual Meeting of the Association for Natural Language Processing*.

Satoshi Sekine, Yoshio Eriguchi, 2000. Results and analyses of irex-ne. In *Proceedings of The Sixth Annual Meeting of the Association for Natural Language Processing*.

TREC, Homepage. *http://trec.nist.gov/*.

TREC_EVAL, FTP site. *ftp://ftp.cs.cornell.edu/pub/smart/trec_eval.v3beta.shar*.

TSC, Homepage. *http://galaga.jaist.ac.jp:8000/tsc/*.