Evaluation of a Practical Interlingua for Task-Oriented Dialogue

Lori Levin, Donna Gates, Alon Lavie, Fabio Pianesi, Dorcas Wallace, Taro Watanabe, Monika Woszczyna
Language Technologies Institute, Carnegie Mellon University and
IRST ITC, Trento, Italy
Internet: lsl@cs.cmu.edu

Abstract

IF (Interchange Format), the interlingua used by the C-STAR consortium, is a speech-act based interlingua for task-oriented dialogue. IF was designed as a practical interlingua that could strike a balance between expressivity and simplicity. If it is too simple, components of meaning will be lost and coverage of unseen data will be low. On the other hand, if it is too complex, it cannot be used with a high degree of consistency by collaborators on different continents. In this paper, we suggest methods for evaluating the coverage of IF and the consistency with which it was used in the C-STAR consortium.

Introduction

IF (Interchange Format) is an interlingua used by the C-STAR consortium[1] for task-oriented dialogues. Because it is used in five different countries for six different languages, it had to achieve a careful balance between being expressive enough and being simple enough to be used consistently. If it was not expressive enough, components of meaning would be lost and coverage of unseen data would be low. On the other hand, if was not simple enough, different system developers would use it inconsistently and the wrong meanings would be translated. IF is described in our previous papers ([PT98, LGLW98, LLW+]).

For this paper, we have proposed methods for evaluating the coverage of IF and the degree to which it can be used consistently across C-STAR sites. Coverage was measured by having human IF specialists annotate unseen data. Consistency was measured by two means. The first was inter-coder agreement among IF specialists at Carnegie Mellon University and ITC-irst (Centro per la ricerca scientifica e tecnologica). The second, less direct method, was a cross-site end-to-end evaluation of English-to-Italian translation where the English-to-IF analysis grammars were written at CMU and IF-to-Italian generation was developed at IRST. If the English and Italian grammar writers did not agree on the meaning of the IF, wrong translations will be produced. In this way, the cross-site evaluation can be an indirect indicator of whether the CMU and IRST IF specialists agreed on the meaning of IF representations. For comparison, we also present within-site end-to-end evaluations of English-to-German, English-to-Japanese, and English-to-IF-to-English, where all of the analysis and generation grammars were written at CMU.

[1] http://www.c-star.org

The Interchange Format

Because we are working with task-oriented dialogues, adequate rendering of the speech act in the target language often overshadows the need for literal translation of the words. IF is therefore based on *domain actions* (DAs), which consist of on speech acts plus domain-specific concepts. An example of a DA is `give-information+price+room` (giving information about the price of a room). DAs are composed from 45 general speech acts (e.g., `acknowledge`, `give-information`, `accept`) and about 96 domain-specific concepts (e.g, `price`, `temporal`, `room`, `flight`, `availability`). In addition to the DA, IF representations can contain arguments such as `room-type`, `destination`, and `price`. There are about 119 argument types.

In the following example, the DA consists of a speaker tag (`a:` for agent), the speech-act `give-information`, and two main concepts, `+price` and `+room`. The DA is followed by a list of arguments: `room-type=` and `price=`. The arguments have values that represent information for the type of room `double` and the cost repre-

Percent Cumulative Coverage	Percent	Count	DA
15.7	15.7	652	acknowledge
19.8	4.1	172	affirm
23.3	3.4	143	thank
26.0	2.7	113	introduce-self
28.0	2.0	85	give-information+price
30.1	2.0	85	greeting
31.9	1.9	78	give-information+temporal
33.7	1.8	75	give-information+numeral
35.5	1.8	73	give-information+price+room
37.2	1.7	70	request-information+payment
38.8	1.6	66	give-information+payment
40.3	1.5	64	give-inform+features+room
41.7	1.4	60	give-inform+availability+room
43.2	1.4	60	accept
44.5	1.3	56	give-information+personal-data
45.8	1.3	52	req-act+reserv+features+room
46.9	1.2	48	req-verif-give-inform+numeral
48.0	1.1	46	offer+help
49.1	1.1	44	apologize
50.1	1.0	42	request-inform+personal-data
...	...	...	...
NA*	5.9	244	no-tag

Figure 1: Coverage of Top 20 DAs and No-tag in development data

Percent Cumulative Coverage	Percent	Count	Speech Act
30.1	30.1	1250	give-information
45.8	15.7	655	acknowledge
57.7	11.9	493	request-information
62.7	5.0	209	request-verification-give-inform
67.6	4.9	203	request-action
71.7	4.1	172	affirm
75.1	3.4	143	thank
77.9	2.7	113	introduce-self
80.2	2.4	98	offer
82.4	2.1	89	accept
84.4	2.0	85	greeting
85.7	1.3	55	suggest
86.8	1.1	44	apologize
87.8	1.0	41	closing
88.5	0.8	32	negate-give-information
89.2	0.6	27	delay-action
89.8	0.6	25	introduce-topic
90.2	0.5	19	please-wait
90.6	0.4	15	reject
91.0	0.4	15	request-suggestion

Figure 2: Coverage of speech-acts in development data

sented with the complex argument `price=` which has its own arguments `quantity=`, `currency=` and `per-unit=`. This IF representation is neutrál between sentences that have different verbs, subjects, and objects such as *A double room costs 150 dollars a night*, *The price of a double room is 150 dollars a night*, and *A double room is 150 dollars a night*.[2]

```
AGENT: ''a double room costs $150 a night.''
a:give-information+price+room
   (room-type=double,
    price=(quantity=150,
           currency=dollar,
           per-unit=night)
```

Coverage and Distribution of Dialogue Acts

In this section, we address the coverage of IF for task-oriented dialogues about travel planning. We want to know whether a very simple interlingua like IF can have good coverage. We are using a rather subjective measure of coverage: IF experts hand-tagged unseen data with IF representations and counted the percentage of utterances to which no IF could be assigned. (When they tagged the unseen data, they were not told that the IF was being tested for coverage. The tagging was done for system development purposes.) Our end-to-end evaluation described in the following sections can be taken as a less subjective measure of coverage. However, the score of an end-to-end evaluation encompasses grammar coverage problems as well as IF coverage problems.

The development portion of the coverage experiment proceeded as follows. Over a period of two years, a database of travel planning dialogues was collected by C-STAR partners in the U.S., Italy, and Korea. The dialogues were role-playing dialogues between a person pretending to be a traveller and a person pretending to be a travel agent. For the English and Italian dialogues, the traveller and agent were talking face-to-face in the same language — both speaking English or both speaking Italian. The Korean dialogues were also role playing dialogues, but one participant was speaking Korean and the other was speaking English. From these dialogues, only the Korean utterances are included in the database. Each utterance in the database is annotated with an English translation and an IF representation. Table 1 summarizes the amount of data in each language. The English, Italian, and Korean data was used for IF development.

The development database contains over 4000 dialogue act units, which are covered by a total of about 542 distinct DAs (346 agent DAs and 278 client DAs). Figures 1 and 2 show the cumulative coverage of the top twenty DA's and speech acts in the development data. Figure 1 also shows the percentage of `no-tag` utterances (the ones we decided not to cover) in the development data. The first column shows the percent of the development data that is covered cumulatively by the DA's or speech acts from the top of the table to the current line. For example, `acknowledge` and `affirm` together account for 19.8 percent of the data. The

[2]When we add anaphora resolution, we will need to know whether a verb (*cost*) or a noun (*price*) was used. This will be an issue our new project, NESPOLE! (`http://nespole.itc.it/`).

Language(s)	Type of Dialogue	Number of DA Units
Development Data:		
English	monolingual	2698
Italian	monolingual	234
Korean-English	bilingual (only Korean utterances are included)	1142
Test Data:		
Japanese-English	bilingual (Japanese and English utterances are included)	6069

Table 1: The IF Database

Percent Cumulative Coverage	Percent	Count	DA
NA*	4.6	263	no-tag
15.6	15.6	885	acknowledge
20.2	4.6	260	thank
23.7	3.5	200	introduce-self
27.0	3.4	191	affirm
29.7	2.7	153	apologize
32.3	2.6	147	greeting
34.6	2.3	128	closing
36.3	1.7	98	give-information+personal-data
38.0	1.7	95	give-information+temporal
39.5	1.6	89	give-information+price
41.1	1.5	88	please-wait
42.5	1.4	82	give-inform+telephone-number
43.8	1.3	75	give-information+features+room
45.0	1.1	65	request-inform+personal-data
46.0	1.0	59	give-inform+temporal+arrival
47.0	1.0	55	accept
48.0	1.0	55	give-inform+availability+room
48.9	1.0	55	give-information+price+room
49.8	0.9	50	verify
50.7	0.9	49	request-inform+temporal+arrival

Figure 3: Coverage of Top 20 DAs and No-tag in test data

Percent Cumulative Coverage	Percent	Count	DA
25.6	25.6	1454	give-information
41.7	16.1	916	acknowledge
53.6	11.9	677	request-information
58.2	4.6	260	thank
62.0	3.7	213	request-verification-give-inform
65.5	3.5	200	introduce-self
68.8	3.4	191	affirm
72.0	3.2	181	request-action
74.8	2.8	159	accept
77.5	2.7	153	apologize
80.1	2.6	147	greeting
82.4	2.3	130	closing
84.4	2.1	117	suggest
86.3	1.8	104	verify-give-information
87.9	1.7	94	offer
89.5	1.5	88	please-wait
90.6	1.1	65	negate-give-information
91.5	0.9	50	verify
92.0	0.5	30	negate
92.5	0.5	26	request-affirmation

Figure 4: Coverage of Top 20 SAs in test data

second column shows the percent of the development data covered by each DA or speech act. The third column shows the number of times each DA or speech act occurs in the development data.

The evaluation portion of the coverage experiment was carried out on 124 dialogues (6069 dialogue act units) that were collected at ATR, Japan. One participant in each dialogue was speaking Japanese and the other was speaking English. Both Japanese and English utterances are included in the data. The 124 Japanese-English dialogues were not examined closely by system developers during IF development. After the IF design was finalized and frozen in Summer 1999, the Japanese-English data was tagged with IFs. No further IF development took place at this point except that values for arguments were added. For example, *Miyako* could be added as a hotel name, but no new speech acts, concepts, or argument types could be added. Sentences were tagged as `no-tag` if the IF did not cover them.

Figures 3 and 4 show the cumulative coverage of the top twenty DAs and speech acts in the Japanese-English data, including the percent of `no-tag` sentences.

Notice that the percentage of `no-tag` was lower in our test data than in our development data. This is because the role playing instructions for the test data were more restrictive than the role playing instructions for the development data. Figures 1 and 3 show that slightly more of the test data is covered by slightly fewer DAs.

Cross-Site Reliability of IF Representations

In this section we attempt to measure how reliably IF is used by researchers at different sites. Recall that one of the design criteria of IF was consistency of use by researchers who are separated by oceans. This criterion limits the complexity of IF. Two measures of consistency are used – inter-coder agreement and a cross-site end-to-end evaluation.

Inter-Coder Agreement: Inter-coder agreement is a direct measure of consistency among

	Percent Agreement
Speech-act	82.14
Dialog-act	65.48
Concept lists	88.00
Argument lists	85.79

Table 2: Inter-coder Agreement between CMU and IRST

C-STAR partners. We used 84 DA units from the Japanese-English data described above. The 84 DA units consisted of some coherent dialogue fragments and and some isolated sentences. The data was coded at CMU and at IRST. We counted agreement on the components of the IF separately. Table 2 shows agreement on speech acts, dialogue acts (speech act plus concepts), concepts, and arguments. The results are reported in Table 2 in terms of percent agreement. Further work might include some other calculation of agreement such as Kappa or precision and recall of the coders against each other. Figure 5 shows a fragment of a dialogue coded by CMU and IRST. The coders disagreed on the IF middle sentence, *I'd like a twin room please*. One coded it as an acceptance of a twin room, the other coded it as a preference for a twin room.

Cross-Site Evaluation: As an approximate and indirect measure of consistency, we have compared intra-site end-to-end evaluation with cross-site end-to-end evaluation. An end-to-end evaluation includes an analyzer, which maps the source language input into IF and a generator, which maps IF into target language sentences. The intra-site evaluation was carried out on English-German, English-Japanese, and English-IF-English translation. The English analyzer and the German, Japanese, and English generators were all written at CMU by IF experts who worked closely with each other. The cross-site evaluation was carried out on English-Italian translation, involving an English analyzer written at CMU and an Italian generator written at IRST. The IF experts at CMU and IRST were in occasional contact with each other by email, and met in person two or three times between 1997 and 1999.

A number of factors contribute to the success of an inter-site evaluation, just one of which is that the sites used IF consistently with each other. Another factor is that the two sites used similar development data and have approximately the same coverage. If the inter-site evaluation results are about as good as the intra-site results, we can conclude that all factors are handled acceptably, including consistency of IF usage. If the inter-site results are worse than the intra-site results, consistency of IF use or some other factor may be to blame. Before conducting this evaluation, we already knew that there was some degree of cross-site consistency in IF usage because we conducted successful inter-continental demos with speech translation and video conferencing in Summer 1999. (The demos and some of the press coverage are reported on the C-STAR web site.) The demos included dialogues in English-Italian, English-German, English-Japanese, English-Korean, and English-French. At a later date, an Italian-Korean demo was produced with no additional work, thus illustrating the well-cited advantage of an interlingual approach in a multi-lingual situation. The end-to-end evaluation reported here goes beyond the demo situation to include data that was unseen by system developers.

Evaluation Data: The Summer 1999 intra-site evaluation was conducted on about 130 utterances from a CMU user study. The traveller was played by a second time user — someone who had participated in one previous user study, but had no other experience with our MT system. The travel agent was played by a system developer. Both people were speaking English, but they were in different rooms, and their utterances were paraphrased using IF. The end-to-end procedure was that (1) an English utterance was spoken and decoded by the JANUS speech recognizer, (2) the output of the recognizer was parsed into an IF representation, and (3) a different English utterance (supposedly with the same meaning) was generated from the IF representation. The speakers had no other means of communication with each other.

In order to evaluate English-German and English-Japanese translation, the IFs of the 130 test sentences were fed into German and Japanese generation components at CMU. The data used in the evaluation was unseen by system developers at the time of the evaluation. For English-Italian translation, the IF representations produced by the English analysis component were sent to IRST to be generated in Italian.

Evaluation Scoring: In order to score the evaluation, input and output sentences were compared by bilingual people, or monolingual people in the case of English-IF-English evaluation. A score of ok is assigned if the target language utterance is comprehensible and no components of meaning are deleted, added, or changed by the translation. A

```
We have singles, and twins and also Japanese rooms available on the eleventh.
CMU  a:give-information+availability+room
(room-type=(single & twin & japanese_style), time=md11)
IRST  a:give-information+availability+room
(room-type=(single & twin & japanese_style), time=md11)

I'd like a twin room, please.
CMU  c:accept+features+room (room-type=twin)
IRST c:give-information+preference+features+room (room-type=twin)

A twin room is fourteen thousand yen.
CMU  a:give-information+price+room
(room-type=twin, price=(currency=yen, quantity=14000))
IRST  a:give-information+price+room
(room-type=twin, price=(currency=yen, quantity=14000))
```

Figure 5: Examples of IF coding from CMU and IRST

	Method	Output Language	OK+Perfect	Perfect	Grader	No. of Graders
1	Recognition only	English	78 %	62 %	CMU	3
2	Transcription	English	74 %	54 %	CMU	3
3	Recognition	English	59 %	42 %	CMU	3
4	Transcription	Japanese	77 %	59 %	CMU	2
5	Recognition	Japanese	62 %	45 %	CMU	2
6	Transcription	German	70 %	39 %	CMU	2
7	Recognition	German	58 %	34 %	CMU	2
8	Transcription	German	67 %	43 %	IRST	2
9	Recognition	German	59 %	36 %	IRST	2
10	Transcription	Italian	73 %	51 %	IRST	6
11	Recognition	Italian	61 %	42 %	IRST	6

Figure 6: Translation Grades for English to English, Japanese, German, and Italian

score of **perfect** is assigned if, in addition to the previous criteria, the translation is fluent in the target language. A score of bad is assigned if the target language sentence is incomprehensible or some element of meaning has been added, deleted, or changed. The evaluation procedure is described in detail in [GLL+96]. In Figure 6, **acceptable** is the sum of **perfect** and **ok** scores. [3]

Figure 6 shows the results of the intra-site and inter-site evaluations. The first row grades the speech recognition output against a human-produced transcript of what was said. This gives us a ceiling for how well we could do if translation were perfect, given speech recognition errors. Rows 2 through 7 show the results of the intra-site evaluation. All analyzers and generators were written at CMU, and the results were graded by CMU researchers. (The German results are a lower than the English and Japanese results because a shorter time was spent on grammar development.) Rows 8 and 9 report on CMU's intra-site evaluation of English-German translation (the same system as in Rows 6 and 7), but the results were graded by researchers at IRST. Comparing Rows 6 and 7 with Rows 8 and 9, we can check that CMU and IRST graders were using roughly the same grading criteria: a difference of up to ten percent among graders is normal in our experience. Rows 10 and 11 show the results of the inter-site English-Italian evaluation. The CMU English analyzer produced IF representations which were sent to IRST and were fed into IRST's Italian generator. The results were graded by IRST researchers.

Conclusions drawn from the inter-site evaluation: Since the inter-site evaluation results are comparable to the intra-site results, we conclude that researchers at IRST and CMU are using IF at least as consistently as researchers within CMU.

Future Plans

In the next phase of C-STAR, we will cover descriptive sentences (e.g., *The castle was built in the thirteenth century and someone was imprisoned in the tower*) as well as task-oriented sentences. Descriptive sentences will be represented

[3] In another paper ([LBL+00]), we describe a task-based evaluation which focuses on success of communicative goals and how long it takes to achieve them.

in a more traditional frame-based interlingua focusing on lexical meaning and grammatical features of the sentences. We are working on disambiguating literal from task-oriented meanings in context. For example *That's great* could be an acceptance (like *I'll take it*) (task oriented) or could just express appreciation. Sentences may also contain a combination of task oriented (e.g., *Can you tell me*) and descriptive (*how long the castle has been standing*) components.

References

[GLL+96] Donna Gates, A. Lavie, L. Levin, A. Waibel, M. Gavaldà, L. Mayfield, M. Woszczyna, and P. Zhan. End-to-End Evaluation in JANUS: A Speech-to-Speech Translation System. In *Proceedings of ECAI-96*, Budapest, Hungary, 1996.

[LBL+00] Lori Levin, Boris Bartlog, Ariadna Font Llitjos, Donna Gates, Alon Lavie, Dorcas Wallace, Taro Watanabe, and Monika Woszczyna. Lessons Learned from a Task-Based Evaluation of Speech-to-Speech MT. In *Proceedings of LREC 2000*, Athens, Greece, June to appear, 2000.

[LGLW98] Lori Levin, D. Gates, A. Lavie, and A. Waibel. An Interlingua Based on Domain Actions for Machine Translation of Task-Oriented Dialogues. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP'98)*, Sydney, Australia, 1998.

[LLW+] Lori Levin, A. Lavie, M. Woszczyna, D. Gates, M. Gavaldà, D. Koll, and A. Waibel. The Janus-III Translation System. *Machine Translation.* To appear.

[PT98] Fabio Pianesi and Lucia Tovena. Using the Interchange Format for Encoding Spoken Dialogue. In *Proceedings of SIG-IL Workshop*, 1998.