D9: Evaluation of the NESPOLE! Showcase-1 System Summary Report

1 Introduction

Over the past 18 months, the NESPOLE! consortium partners have developed a fully functional showcase of the NESPOLE! system within the domain of travel and tourism, and have significantly improved system performance and usability based on a series of studies and evaluations with real users. A full multi-site, multi-lingual end-to-end evaluation of the system was conducted in the last two months of 2001. In this report, we describe our evaluation methodology and report the results and our experience from this full scale evaluation of the NESPOLE! system.

In the first showcase described here, the scenario is the following: a client user is browsing through the web-pages of APT – the tourism bureau of the province of Trentino in Italy – in search of winter-sport tour-packages in the Trentino region. If more detailed information is desired, the client can click on a dedicated "button" within the web-page in order to establish a video-conferencing connection to a human agent located at APT. The client is then presented with an interface consisting primarily of a standard video-conferencing application window and a shared whiteboard application. Using this interface, the client can carry on a conversation with the agent, where the NESPOLE! server provides two-way speech-to-speech translation between the parties. In the current setup, the agent speaks Italian, while the client can speak English, French or German.

2 Evaluation Methodology

In December 2001, we conducted a large scale multi-lingual end-to-end translation evaluation of the NESPOLE! first-showcase system. For each of the three language pairs (English-Italian, German-Italian and French-Italian), four previously unseen test dialogues were used to evaluate the performance of the translation system. The dialogues included two scenarios: one covering winter ski vacations, the other about summer resorts. One or two of the dialogues for each language contained multi-modal expressions. The test data included a mixture of dialogues that were collected mono-lingually prior to system development (both client and agent spoke the same language), and data collected bilingually (during the July 2001 MM experiment), using the actual translation system. This mixture of data conditions was intended primarily for comprehensiveness and not for comparison of the different conditions.

We performed an extensive suite of evaluations on the above data which we describe below. The evaluations were for the most part end-to-end, from input to output, not assessing individual modules or components. The Speech Recognition modules were also evaluated in isolation, using standard word error rate (WER) metrics, in order to allow us to assess their performance effect on the overall translation performance. We performed both mono-lingual evaluation (where generated output language was the same as the input language), as well as cross-lingual evaluation. For cross-lingual evaluations, translation from English German and French to Italian was evaluated

Language	WARs	SR Graded (% Acc)
English	61.9%	66.0%
German	63.5%	68.0%
French	71.2%	65.0%
Italian	76.5%	70.6%

Table 1: *Speech Recognition Word Accuracy Rates and Results of Human Grading (Percent Acceptable) of Recognition Output as a Paraphrase*

Language	Transcribed	Speech Rec.
English-to-English	58%	45%
German-to-German	46%	40%
French-to-French	54%	41%
Italian-to-Italian	61%	48%

Table 2: *Monolingual End-to-End Translation Results (Percent Acceptable) on Transcribed and Speech Recognized Input*

on client utterances, and translation from Italian to each of the three languages was evaluated on agent utterances. We evaluated on both manually transcribed input as well as on actual speech-recognition of the original audio. We also graded the speech recognized output as a "paraphrase" of the transcriptions, to measure the levels of semantic loss of information due to recognition errors. Speech recognition word accuracies and the results of speech graded as a paraphrase appear in Table 1. Translations were graded by multiple human graders at the level of Semantic Dialogue Units (SDUs). For each data set, one grader first manually segmented each utterance into SDUs. All graders then used this segmentation in order to assign scores for each SDU present in the utterance. We followed the three-point grading scheme previously developed for the C-STAR consortium, as described in [1]. Each SDU is graded as either "Perfect" (meaning translated correctly and output is fluent), "OK" (meaning is translated reasonably correct but output may be disfluent), or "Bad" (meaning not properly translated). We calculate the percent of SDUs that are graded with each of the above categories. "Perfect" and "OK" percentages are also summed together into a category of "Acceptable" translations. Average percentages are calculated for each dialogue, each grader, and separately for client and agent utterances. We then calculated combined averages for all graders and for all dialogues for each language pair.

2.1 Performance Results

Table 2 shows the results of the monolingual end-to-end translation for the four languages, and Table 3 shows the results of the cross-lingual evaluations. The results indicate acceptable translations in the range of 27–43% of SDUs (interlingua units) with speech recognized inputs. While this level of translation accuracy cannot be considered impressive, our user studies and system demonstrations indicate that it is already sufficient for achieving effective communication with real users. We anticipate performance levels will reach "Acceptable" translation in the range of 60–70% of SDUs within the next year of the project.

Language	Transcribed	Speech Rec.
English-to-Italian	55%	43%
German-to-Italian	32%	27%
French-to-Italian	44%	34%
Italian-to-English	47%	37%
Italian-to-German	47%	31%
Italian-to-French	40%	27%

Table 3: *Cross-lingual End-to-End Translation Results (Percent Acceptable) on Transcribed and Speech Recognized Input*

3 Analysis of the Evaluation Scheme

We are currently in the process of several advanced investigations into issues related to our evaluation methodology, based on our experience from the NESPOLE! evaluation reported earlier. These issues will be studied in the last year of the project, and the conclusions will be taken into account in the evaluation of the second showcase at the end of the project. Three main issues are under investigation:

Three-way Versus Binary Grading: The current evaluation scheme requires graders to assign one of three possible grades (Perfect/OK/Bad) in a single pass. The intention behind the three categories, however, is that we are applying two types of judgements: (1) is the meaning of the original SDU preserved in translation? (2) is the translation output fluent and grammatical? If the answer to the first judgement is negative, the grade assigned is “Bad” (regardless of the answer to the second question). The second question distinguishes between the case of a “Perfect” versus and “OK” grade. An alternative way to assign the same judgements is therefore to sequentially pose two binary questions to the graders. We are in the process of conducting an experiment to assess whether there are any significant differences between these two methods of scoring.

Averaging Scores Versus Majority Votes: Our custom in the past, when grading with multiple human subjects, has been to calculate percentages for each of the grade categories (Perfect/OK/Bad) for each subject separately, and then to compute a simple arithmetic average of all the human subjects. In the current NESPOLE! evaluation, we experimented for the first time with also calculating majority votes. Under this scheme, for each SDU, we assign a grade category based on majority (when a majority exists). In cases where there was no majority, or where there was at least one vote for both “Perfect” and “Bad”, the SDU was excluded from the scoring. Majority votes for the current NESPOLE! evaluation were calculated for the Italian and French evaluations, and are included in the detailed tables of results in Appendix-A. A more detailed comparison of the two evaluation methods will be investigated in the next year of the project.

Intercoder and Intracoder Agreement: To establish the stability and coherence of our evaluation scheme, it is important to have a good measure of how well different human graders agree on scoring the same output, and also how consistent the graders are over time. The experiment we are currently conducting is collecting the necessary data to evaluate both of these questions. We will use standard measures from the literature to quantify agreement: the Kappa coefficient, and confusion matrices.

Appendix-A: Detailed Evaluation Result Tables

3.1 Italian Evaluations

```
* Italian Test Dialogues
  ----------------------
  4 dialogues

  a1      = ita060co1     ( 72  SDUs)     (33 utts)
  a2      = ita806co1     ( 20  SDUs)     (10 utts)
  amm     = ita911co1     ( 39  SDUs)     (25 utts)
  c       = ita024co3     ( 69  SDUs)     (22 utts)
  ------------------------------------------------
  ALL     = total         (200  SDUs)     (90 utts)

* Italian SR Accuracy
  -------------------
  a2   =      76.43%
  amm  =      76.92%

* HYPO Graded
  -----------
       G1       G2       G3     ALL   | WA(%)
---------------------------------------------
 a2   60(45)   70(45)  65(40)  65(45) | 76.4
---------------------------------------------
amm   77(64)   74(59)  74(56)  72(59) | 76.9
---------------------------------------------
ALL   71(58)   73(54)  71(51)  70(54) | 76.8
---------------------------------------------

* ITA-to-ITA: SLT-TCT
  -------------------
       G1       G2       G3       AVER     MAJ
--------------------------------------------------
a1     71(38)   78(74)   62(45)   70(52)   69(50)
--------------------------------------------------
a2     42(26)   55(35)   55(35)   51(32)   44(28)
--------------------------------------------------
amm    68(35)   83(73)   60(55)   70(54)   68(55)
--------------------------------------------------
c      54(30)   49(46)   51(38)   51(38)   58(42)
==================================================
Aver   59(32)   66(57)   57(43)   61(44)   60(44)

* ITA-to-ITA: SLT-REC
  -------------------
       G1       G2       G3       AVER     MAJ
--------------------------------------------------
a2     30(20)   38(29)   35(20)   34(23)   33(22)
--------------------------------------------------
amm    56(26)   75(65)   55(35)   62(42)   55(37)
==================================================
Aver   43(23)   57(47)   45(28)   48(33)   44(30)
```

* ITA-to-ENG : SLT-TCT

	G1	G2	G3	AVER	MAJ
a1	61(46)	69(50)	51(36)	60(44)	56(46)
a2	40(30)	30(15)	35(25)	35(23)	37(26)
amm	54(44)	62(36)	51(36)	56(39)	54(41)
c	39(30)	43(26)	33(27)	38(28)	36(26)
Aver	49(38)	51(32)	43(31)	47(34)	46(35)

* ITA-to-ENG : SLT-REC

	G1	G2	G3	AVER	MAJ
a2	26(21)	24(10)	20(20)	23(17)	27(27)
amm	45(28)	65(40)	41(31)	50(33)	43(27)
Aver	36(24)	45(25)	31(26)	37(25)	35(27)

* ITA-to-GER: SLT-TCT

	G1	G2	G3	AVER	MAJ
a1	58(33)	58(31)	55(41)	57(35)	61(41)
a2	50(20)	37(16)	30(20)	39(19)	33(11)
amm	54(23)	46(23)	49(26)	50(24)	47(18)
c	42(25)	40(24)	47(36)	43(28)	39(29)
Aver	51(25)	45(24)	45(31)	47(27)	45(25)

* ITA-to-GER: SLT-REC

	G1	G2	G3	AVER	MAJ
a2	30(15)	17(11)	20(15)	22(14)	13(6)
amm	47(16)	42(16)	29(21)	39(17)	26(11)
Aver	39(16)	30(14)	25(18)	31(16)	20(9)

3.2 English Evaluations

```
* English Test Dialogues
  ----------------------
4 dialogues

a1    = e025ap   ( 46 SDUs)  ( 27 utts)
a2    = e039ap   (123 SDUs)  ( 37 utts)
amm   = e011yp   ( 54 SDUs)  ( 39 utts)
cmm   = e827cy   (109 SDUs)  ( 48 utts)
----------------------------------------
ALL   = total    (332 SDUs)  (151 utts)


* English SR Accuracy
  -------------------
 Speaker      % Accuracy
 ------------------------
 e025ap         68.6
 e039ap         39.5
 e011yp         83.1
 e827cy         71.0
 ------------------------
 Average        61.9

* HYPO Graded
  -----------
       G1       G2       G3      ALL   | WA
-------------------------------------------
 a1   76(65)   74(61)  65(52)  72(59) | 68
-------------------------------------------
 a2   55(39)   43(32)  50(35)  50(35) | 39
-------------------------------------------
amm   91(89)   93(85)  91(78)  91(84) | 84
-------------------------------------------
cmm   71(63)   65(59)  69(56)  68(59) | 70
-------------------------------------------
ALL   69(59)   63(54)  65(51)  66(56) | 61
-------------------------------------------

* Eng-to-Eng: SLT-TCT
  --------------------
        G1       G2       G3       ALL
----------------------------------------
 a1   74(70)   76(54)    67(41)   72(55)
----------------------------------------
 a2   62(46)   45(40)    46(32)   51(39)
----------------------------------------
amm   74(57)   67(54)    61(48)   67(53)
----------------------------------------
cmm   65(49)   40(31)    51(31)   52(37)
----------------------------------------
ALL   67(52)   51(41)    53(35)   58(43)
----------------------------------------
```

* Eng-to-Eng: SLT-REC

	G1	G2	G3	ALL
a1	58(50)	52(33)	43(24)	51(36)
a2	41(27)	29(23)	33(21)	34(23)
amm	69(57)	70(63)	70(41)	70(54)
cmm	50(39)	32(26)	41(21)	41(29)
ALL	51(39)	40(32)	43(25)	45(32)

* English-to-Italian

	a1	a2	amm	cmm	ALL
TCT	77(52)	48(36)	67(45)	59(31)	55(38)
REC	57(39)	29(19)	69(44)	39(24)	43(27)

3.3 German Evaluations

* German Test Dialogues

4 dialogues

a1	= g047ak	(46 SDUs)	(23 utts)
a2	= g051ak	(174 SDUs)	(59 utts)
amm	= g006yk	(108 SDUs)	(70 utts)
c1	= g034ck	(314 SDUs)	(98 utts)
All	= total	(644 SDUs)	(350 utts)

* German SR Accuracy

Speaker	% Accuracy
g006	42.69
g034	66.32
g047	78.67
g051	69.43
Average	63.52

* Graded HYPO

	ALL	G1	G2	G3	G4	G5	WA(%)
ALL	68(51)	57(50)	59(50)	64(48)	86(65)	82(52)	63.5
a1	87(75)	76(72)	85(76)	83(74)	96(80)	96(74)	78.7
a2	80(64)	68(58)	70(61)	76(57)	93(79)	93(62)	69.4
amm	47(29)	36(30)	40(30)	39(15)	69(42)	51(28)	42.7
c1	69(53)	55(50)	56(48)	64(50)	86(62)	85(52)	66.3

* Ger-to-Ger: SLT-TCT

	ALL	G1	G2	G3	G4	G5
ALL	46(20)	28(23)	24 (6)	39 (7)	79(39)	61(24)
a1	64(34)	50(41)	48(11)	67(11)	78(61)	78(43)
a2	50(22)	34(30)	29 (5)	43 (6)	73(47)	70(20)
amm	46(23)	32(24)	30(16)	39(15)	77(40)	54(22)
c1	41(16)	20(17)	17 (4)	32 (5)	83(31)	55(23)

* Ger-to-Ger: SLT-REC

	ALL	G1	G2	G3	G4	G5
ALL	40(18)	26(23)	21 (5)	32 (5)	65(32)	57(21)
a1	59(31)	50(41)	46 (9)	60(11)	70(54)	72(39)
a2	49(21)	37(32)	25 (4)	39 (5)	81(42)	61(22)
amm	32(14)	20(17)	19(10)	28 (8)	56(22)	38(14)
c1	34(14)	18(17)	16 (3)	24 (4)	59(26)	59(20)

* Ger-to-Ita:

	G1	G2	G3	All
SLT-TCT	31 (7)	38 (9)	30 (24)	32 (13)
SLT-REC	26 (4)	32 (6)	26 (22)	27 (11)

3.4 French Evaluations

* French Test Dialogues

4 dialogues

A1	= srA1	(109 SDUs)	(60 utts)
A2	= lbA2	(139 SDUs)	(74 utts)
C3	= srC3	(101 SDUs)	(64 utts)
C4	= lbC4	(78 SDUs)	(37 utts)
All	= total	(427 SDUs)	(235 utts)

* French SR Accuracy

Dialogue	% Accuracy
A1	80.1%
A2	57.4%
C3	74.4%
C4	81.8%
Average	71.2%

* Graded HYPO

	G1	G2	G3	AVER	MAJ
A1	74.5(70.0)	70.9(68.2)	73.6(70.0)	73.0(69.4)	72.5(69.7)
A2	49.3(45.0)	46.8(40.4)	47.1(42.1)	47.7(42.5)	46.0(41.0)
C3	68.0(61.2)	64.1(59.2)	69.9(60.2)	67.3(60.2)	68.3(61.4)
C4	81.0(78.5)	80.0(78.8)	82.5(77.5)	81.2(78.3)	76.9(75.6)
Aver	66.0(61.3)	63.1(59.0)	65.8(60.0)	65.0(60.1)	63.7(59.5)

* FRE-to-FRE: SLT-TCT

	G1	G2	G3	AVER	MAJ
A1	67.7(48.6)	63.6(50.9)	67.6(55.9)	66.3(51.8)	61.5(52.3)
A2	44.7(34.0)	44.7(38.3)	47.9(36.6)	45.8(36.3)	40.3(36.0)
C3	49.5(36.9)	45.6(40.8)	49.5(41.7)	48.2(39.8)	48.5(40.6)
C4	64.6(49.4)	56.4(52.6)	62.0(51.9)	61.0(51.3)	53.8(47.4)
Aver	55.3(41.2)	51.8(44.7)	55.9(45.5)	54.3(43.8)	50.7(43.3)

* FRE-to-FRE: SLT-REC

	G1	G2	G3	AVER	MAJ
A1	49.1(31.3)	54.6(33.6)	51.8(34.5)	51.8(33.1)	49.5(31.2)
A2	26.4(15.0)	28.2(18.3)	27.0(19.9)	27.2(17.7)	24.5(14.4)
C3	37.3(24.5)	40.8(29.1)	35.9(30.1)	38.0(27.9)	36.6(27.7)
C4	52.5(32.5)	57.7(42.3)	51.3(46.3)	53.8(40.4)	46.2(37.2)
Aver	39.6(24.7)	43.2(29.1)	39.9(30.9)	40.9(28.2)	37.7(26.0)

* FRE-to-ITA: SLT-TCT

	G1	G2	G3	AVER	MAJ
A1	56.0(40.4)	51.4(38.5)	59.8(47.7)	55.7(42.2)	48.6(37.6)
A2	34.8(24.1)	25.0(21.3)	37.9(30.7)	32.6(25.4)	27.3(23.7)
C3	40.8(31.1)	34.3(30.4)	43.1(37.3)	39.4(32.9)	39.6(34.7)
C4	50.0(36.3)	50.7(42.5)	56.3(46.3)	52.3(41.7)	41.0(34.6)
Aver	44.3(32.1)	38.6(31.7)	48.0(39.4)	43.6(34.4)	38.2(31.9)

* FRE-to-ITA: SLT-REC

	G1	G2	G3	AVER	MAJ
A1	39.1(27.3)	43.5(35.2)	39.1(29.1)	40.6(30.5)	37.6(28.4)
A2	23.4(15.6)	20.1(11.5)	22.9(19.3)	22.1(15.5)	18.7(15.1)
C3	32.0(24.3)	30.7(23.8)	33.0(27.2)	31.9(25.1)	30.7 (25.7)
C4	46.3(36.3)	46.8(36.7)	51.3(41.3)	48.1(38.1)	42.3(35.9)
Aver	33.6 24.4)	33.5(25.1)	34.6(27.7)	33.9(25.7)	30.7(24.8)

References

[1] Lori Levin, Donna Gates, Fabio Pianesi, Donna Wallace, Takeshi Watanabe, and Monika Woszczyna. Evaluation of a Practical Interlingua for Task-Oriented Dialogues. In *Proceedings NAACL-2000 Workshop On Interlinguas and Interlingual Approaches*, Seattle, WA, 2000. AMTA.