APPENDIX B

Results of the Multiple Language Question Answering Track (QA@CLEF)

Prepared by:

Alessandro Vallin[†] and Jesús Herrera[‡]

[†] ITC-Irst, Trento, Italy, *vallin@itc.it*
[‡] Dpto. Lenguajes y Sistemas Informáticos, UNED, Madrid, Spain, *jesus.herrera@lsi.uned.es*

List of Run Characteristics

The following table lists the 18 participating teams and the 49 runs submitted at the CLEF-2004 Question Answering Track.

GROUP	COUNTRY	TASKS								RUNS
		monolingual						bilingual	pilot ES	
		DE	ES	FR	IT	NL	PT			
Bulgarian Academy of Sciences	Bulgaria							BG=>EN		bgas041bgen
U. Da Coruna	Spain		●							cole041eses
DAEDALUS (*)	Spain		●							mira041eses
DFKI	Germany	●						DE=>EN		dfki041dede dfki041deen
U.Helsinki	Finland							FI=>EN		hels041fien
U.Edinburgh	UK							DE=>EN FR=>EN		edin041deen edin042deen edin041fren edin042fren
ILC-CNR	Italy				●					ILCP-QA-ITIT
Inst. Nac. Astrofisica, Optica y Electron.	Mexico		●							inao041eses inao042eses
ITC-irst, TCC Division	Italy				●			IT=>EN		irst041itit irst042itit irst041iten irst042iten
Linguateca, Sintef	Norway						●			sfnx041ptpt sfnx042ptpt
LIMSI-CNRS	France							FR=>EN		lire041fren lire042fren
U.Politecnica Catalunya	Spain		●							talp041eses talp042eses
U. Alicante	Spain		●						●	aliv041eses aliv042eses alivpilot
U. Amsterdam	Netherlands					●		EN=>NL		uams041nlnl uams042nlnl uams041ennl
U.Evora	Portugal						●			PTUE041ptpt
U.Hagen	Germany	●								FUHA041dede
U.Limerick	Ireland							FR=>EN		dltg041fren dltg042fren
U.Neuchatel	Switzerland			●				BG=>FR DE=>FR EN=>FR ES=>FR IT=>FR NL=>FR PT=>FR		gine041frfr gine042frfr gine041bgfr gine042bgfr gine041defr gine042defr gine041enfr gine042enfr gine041esfr gine042esfr gine041itfr gine042itfr gine041nlfr gine042nlfr gine041ptfr gine042ptfr

(*) The DAEDALUS group submitted the results after the scheduled deadline.

Results for Main Tasks

In the following six pages the results for the main QA tasks are given. They are divided according to target languages, so that there is a separate table per language. Several tasks can be grouped in the same target language.

Each table provides the following information:

- the **name of the submitted run**;
- the **task** in which the group participated;
- the **number of answers** contained in each submission (divided into **Right**, **Wrong**, **ineXact** and **Unsupported**). In all the tasks there were 200 questions and systems were allowed to return just one response per question. Nevertheless, some runs count less than 200 answers, because some questions that contained mistakes were discarded;
- the **overall accuracy** of each run (i.e. the percentage of Right answers);
- the **accuracy over the Factoid questions**;
- the **accuracy over the Definition questions** (test sets contained around 20 of them);
- the systems' **Precision** and **Recall** in recognising the questions that did not have any answer (the correct answer-string was "**NIL**");
- the **Confidence-weighted Score**, which takes into account the systems' ability to rank the answers according to confidence. This additional measure ranges between 0 (no correct response at all) and 1 (all the answers are correct and the system is always confident about them). Since the confidence value was not mandatory, the Confidence-weighted Score was not computed for all the runs.

German (DE) as target language:

Run Name	Task	# Answers	# Right	# Wrong	# ineXact	# Unsupported	Overall Accuracy %	Accuracy over F %	Accuracy over D %	NIL Accuracy		Confidence weighted Score
										Precision	Recall	
dfki041dede	DE=>DE	197	50	143	1	3	25.3	28.25	0	0.13	0.85	/
FUHA041dede	DE=>DE	197	67	128	2	0	34	31.64	55	0.13	1	0.333

English (EN) as target language:

Run Name	Task	# Answers	# Right	# Wrong	# ineXact	# Unsupported	Overall Accuracy %	Accuracy over F %	Accuracy over D %	NIL Accuracy		Confidence weighted Score
										Precision	Recall	
bgas041bgen	BG=>EN	200	26	168	5	1	13	11.6	25	0.13	0.4	0.056
dfki041deen	DE=>EN	200	47	151	0	2	23.5	23.8	20	0.1	0.75	0.177
dltg041fren	FR=>EN	200	38	155	7	0	19	17.7	30	0.17	0.55	/
dltg042fren	FR=>EN	200	29	164	7	0	14.5	12.7	30	0.14	0.45	/
edin041deen	DE=>EN	200	28	166	5	1	14	13.3	20	0.14	0.35	0.049
edin041fren	FR=>EN	200	33	161	6	0	16.5	17.7	5	0.15	0.55	0.056
edin042deen	DE=>EN	200	34	159	7	0	17	16.1	25	0.14	0.35	0.052
edin042fren	FR=>EN	200	40	153	7	0	20	20.5	15	0.15	0.55	0.058
hels041fien	FI=>EN	193	21	171	1	0	10.8	11.5	5	0.1	0.85	0.046
irst041iten	IT=>EN	200	45	146	6	3	22.5	22.2	25	0.24	0.3	0.121
irst042iten	IT=>EN	200	35	158	5	2	17.5	16.6	25	0.24	0.3	0.075
lire041fren	FR=>EN	200	22	172	6	0	11	10	20	0.05	0.05	0.032
lire042fren	FR=>EN	200	39	155	6	0	19.5	20	15	0	0	0.075

Spanish (ES) as target language:

Run Name	Task	# Answers	# Right	# Wrong	# ineXact	# Unsupported	Overall Accuracy %	Accuracy over F %	Accuracy over D %	NIL Accuracy		Confidence weighted Score
										Precision	Recall	
aliv041eses	ES=>ES	200	63	130	5	2	31.5	30.5	40	0.17	0.35	0.121
aliv042eses	ES=>ES	200	65	129	4	2	32.5	31.1	45	0.17	0.35	0.144
cole041eses	ES=>ES	200	22	178	0	0	11	11.6	5	0.1	1	/
inao041eses	ES=>ES	200	45	145	5	5	22.5	19.44	50	0.19	0.5	/
inao042eses	ES=>ES	200	37	152	6	5	18.5	17.78	25	0.21	0.5	/
mira041eses	ES=>ES	200	18	174	7	1	9	10	0	0.14	0.55	/
talp041eses	ES=>ES	200	48	150	1	1	24	18.8	70	0.19	0.5	0.087
talp042eses	ES=>ES	200	52	143	3	2	26	21.1	70	0.2	0.55	0.102

French (FR) as target language:

Run Name	Task	# Answers	# Right	# Wrong	# ineXact	# Unsupported	Overall Accuracy %	Accuracy over F %	Accuracy over D %	NIL Accuracy		Confidence weighted Score
										Precision	Recall	
gine041bgfr	BG=>FR	200	13	182	5	0	6.5	6.6	5	0.1	0.5	0.051
gine041defr	DE=>FR	200	27	162	11	0	13.5	13.8	10	0.15	0.2	0.071
gine041enfr	EN=>FR	200	16	171	13	0	8	8.3	5	0.05	0.1	0.031
gine041esfr	ES=>FR	200	25	166	9	0	12.5	13.8	0	0.12	0.15	0.054
gine041frfr	FR=>FR	200	26	160	14	0	13	13.8	5	0	0	0.046
gine041itfr	IT=>FR	200	23	166	11	0	11.5	12.2	5	0.15	0.3	0.047
gine041nlfr	NL=>FR	200	17	171	12	0	8.5	8.8	5	0.12	0.2	0.041
gine041ptfr	PT=>FR	200	22	170	8	0	11	11.1	10	0.11	0.15	0.041
gine042bgfr	BG=>FR	200	13	180	7	0	6.5	6.1	10	0.1	0.35	0.038
gine042defr	DE=>FR	200	32	155	13	0	16	15	25	0.23	0.2	0.087
gine042enfr	EN=>FR	200	25	165	10	0	12.5	11.6	20	0.06	0.1	0.048
gine042esfr	ES=>FR	200	30	164	6	0	15	15.5	10	0.11	0.1	0.063
gine042frfr	FR=>FR	200	42	147	11	0	21	20.5	25	0.09	0.05	0.095
gine042itfr	IT=>FR	200	27	165	8	0	13.5	14.4	5	0.14	0.3	0.052
gine042nlfr	NL=>FR	200	26	158	16	0	13	12.2	20	0.14	0.2	0.06
gine042ptfr	PT=>FR	200	25	166	9	0	12.5	11.6	20	0.1	0.15	0.05

Italian (IT) as target language:

Run Name	Task	# Answers	# Right	# Wrong	# ineXact	# Unsupported	Overall Accuracy %	Accuracy over F %	Accuracy over D %	NIL Accuracy		Confidence weighted Score
										Precision	Recall	
ILCP-QA-ITIT	IT=>IT	200	51	117	29	3	25.5	22.7	50	0.62	0.5	/
irst041itit	IT=>IT	200	56	131	11	2	28	26.6	40	0.27	0.3	0.155
irst042itit	IT=>IT	200	44	147	9	0	22	20	40	0.66	0.2	0.107

Dutch (NL) as target language:

Run Name	Task	# Answers	# Right	# Wrong	# ineXact	# Unsupported	Overall Accuracy %	Accuracy over F %	Accuracy over D %	NIL Accuracy		Confidence weighted Score
										Precision	Recall	
uams041ennl	EN=>NL	200	70	122	7	1	35	31	65.2	0	0	0.222
uams041nlnl	NL=>NL	200	88	98	10	4	44	42.3	56.5	0	0	0.284
uams042nlnl	NL=>NL	200	91	97	10	2	45.5	45.2	47.8	0.55	0.25	0.326

Portuguese (PT) as target language:

Run Name	Task	# Answers	# Right	# Wrong	# ineXact	# Unsupported	Overall Accuracy %	Accuracy over F %	Accuracy over D %	NIL Accuracy		Confidence weighted Score
										Precision	Recall	
PTUE041ptpt	PT=>PT	199	56	125	18	0	28.1	28.5	25.8	0.14	0.9	0.243
sfnx041ptpt	PT=>PT	199	22	165	8	4	11	11.9	6.4	0.13	0.7	/
sfnx042ptpt	PT=>PT	199	30	154	10	5	16	11.3	9.6	0.16	0.6	/

Spanish Pilot Task

An additional pilot task was set up only for Spanish. Differently from the main tasks, list questions and questions that required more sophisticated temporal reasoning were proposed.
The following table describes the results of the run *alivpilot*, submitted by the University of Alicante, that was the only participating team. Results have been grouped by type of question (definition, factoid, list, temporally restricted by date, temporally restricted by event and temporally restricted by period).
In addition, a couple of the posed questions had no answer in the corpus (NIL) but the system did not recognise them.

The table provides the following information:
- the **number of questions**;
- the **number of known distinct answers**, i.e., the number of different and correct answers retrieved by the University of Alicante system in its exercise and by humans during the pre-assessment process;
- the **number of given answers**;
- the **number of questions with at least 1 correct answer**, i.e., questions with at least 1 answer assessed as **Right**;
- the **number of given correct answers**;
- the system's **recall** in recognising correct answers, i.e., the ratio between the number of given correct answers and the number of known distinct answers;
- the system's **precision** in recognising correct answers, i.e., the ratio between the number of given correct answers and the number of given answers;
- the ***K*-measure**[1] value; this metrics ranges in [-1, 1] and rewards systems that:
  - answer as many questions as possible,
  - give as many different right answers for each question as possible,
  - give the smaller number of wrong answers to each question,
  - assign higher values of the score to right answers,
  - assign lower values of the score to wrong answers,
  - give answer to the questions having less known answers;
- the **correlation coefficient (r)** between the confidence score and human assessment; human assessment equals 1 when an answer is assessed as **Right** and 0 otherwise; **r** gives an idea about the quality of the system's self-scoring.

		# questions	# known distinct answers	# given answers	# questions with at least 1 correct answer	# given correct answers	recall	precision	*K*	r
Definition		2	3	2	0 (0%)	0	0%	0%	0	N/A †
Factoid		18	26	42	4 (22.2%)	5	19.2%	11.9%	-0.029	-0.089
List		20	191	55	4 (20%)	6	3.1%	10.9%	-0.07	0.284
Temp.	**Date**	20	20	30	2 (10%)	2	10%	6.6%	-0.019	N/A
	Event	20	20	42	2 (10%)	2	10%	4.7%	-0.024	0.255
	Period	20	20	29	3 (15%)	3	15%	10.3%	-0.003	0.648
Total		100	280	200	15 (15%)	18	6.4%	9%	-0.086	0.246

† **r** is Not Available because 0 was given for every component of any variable.

[1] *K*-measure is defined in: J. Herrera, A. Peñas, and F. Verdejo. Question Answering Pilot Task at CLEF 2004. In *Proceedings of the CLEF 2004 Workshop*, Bath, United Kingdom, September 2004.