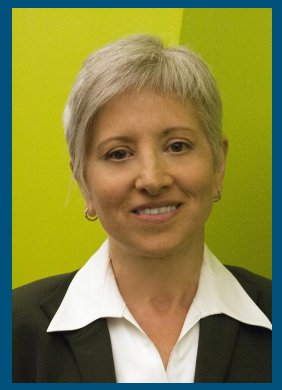
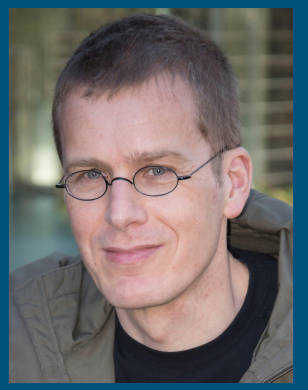
CODI-CRAC 2022: Shared-Task

Anaphora, Bridging, and Discourse Deixis in Dialogue

Contact: sharedtask-codicrac-coling2022@googlegroups.com




Juntao Yu, Sopan Khosla, Ramesh Manuvinakurike, Lori Levin, Vincent Ng, Massimo Poesio, Michael Strube, Carolyn Rose

Anaphoric Relations

- Anaphoric Identity **(OntoNotes, GAP, LitBank, ARRAU, SemEval 2010, GUM,…)**
- Bridging **(ARRAU, ISNotes, BASHI, Prague Dependency Treebank,...)**
- Discourse Deixis/Abstract Anaphora **(ARRAU, ASN, ...)**

Anaphoric Relations

- Anaphoric Identity **(OntoNotes, GAP, LitBank, ARRAU, SemEval 2010, GUM,…)**
- Bridging **(ARRAU, ISNotes, BASHI, Prague Dependency Treebank,...)**
- Discourse Deixis/Abstract Anaphora **(ARRAU, ASN, ...)**
- **Not many state-of-the-art dialogue datasets!**

The first edition:

- Dialogues + Three Anaphoric Relations
  - Anaphoric Identity, Bridging, Discourse Deixis

The second edition

Same tasks: *anaphoric identity, bridging references, discourse deixis*

Same Genre: *Conversation*

Same Opportunities for interaction between communities: *Discourse and Dialogue*

New Computational techniques: *from no-resource to low-resource (additional in-domain training data provided)*

New *data set (with a focus on improve the quality of the annotation)*

New *stages for bridging and discourse deixis with gold anaphors be provided.*

CODI-CRAC 2022: Data

- Out-of-domain training-data
  - Non-conversational: ARRAU_PEAR, ARRAU_RST, ARRAU_GNOME, etc.
  - Conversational: ARRAU_Trains93
- In-domain training-data
  - AMI, Switchboard, Light, Preservation
- Participants were encouraged to use other datasets!

CODI-CRAC 2022: Data

- Newly annotated test-set from multiple domains
- Revised train, dev sets

		Docs	Tokens	Markables	DO	Bridging	Disc. Deix
LIGHT	train	20	11495	3907	2132	381	72
	dev	21	11824	3941	2181	424	84
	test	38	22017	7330	3770	812	128
AMI	train	7	33741	8918	4579	853	230
	dev	3	18260	4870	2350	638	118
	test	3	16562	3990	2007	432	118
PERSUASION	train	21	9185	2743	1242	248	95
	dev	27	12198	3697	1715	316	133
	test	33	14719	4233	2111	304	105
SWITCHBOARD	train	11	14992	4024	1679	589	128
	dev	22	35027	9392	3991	1165	265
	test	12	14605	3888	1606	464	107
Total		218	214625	60933	29363	6626	1583

Table 1: Statistics about the CODI-CRAC 2022 corpus (new datasets only)

CODI-CRAC 2022: Data

- 21K tokens for train, dev, and test
- Annotation Details
  - ARRAU revised annotation scheme
  - Work split between QMUL and CMU
  - Annotated with MMAX2
    - Müller and Strube, 2006
  - Universal Anaphora Format
- Availability
  - ARRAU & SWBD. distributed by LDC
  - AMI, LIGHT & PERS. on Shared Task site

		Docs	Tokens	Markables	DO	Bridging	Disc. Deix
LIGHT	train	20	11495	3907	2132	381	72
	dev	21	11824	3941	2181	424	84
	test	38	22017	7330	3770	812	128
AMI	train	7	33741	8918	4579	853	230
	dev	3	18260	4870	2350	638	118
	test	3	16562	3990	2007	432	118
PERSUASION	train	21	9185	2743	1242	248	95
	dev	27	12198	3697	1715	316	133
	test	33	14719	4233	2111	304	105
SWITCHBOARD	train	11	14992	4024	1679	589	128
	dev	22	35027	9392	3991	1165	265
	test	12	14605	3888	1606	464	107
Total		218	214625	60933	29363	6626	1583

Table 1: Statistics about the CODI-CRAC 2022 corpus (new datasets only)

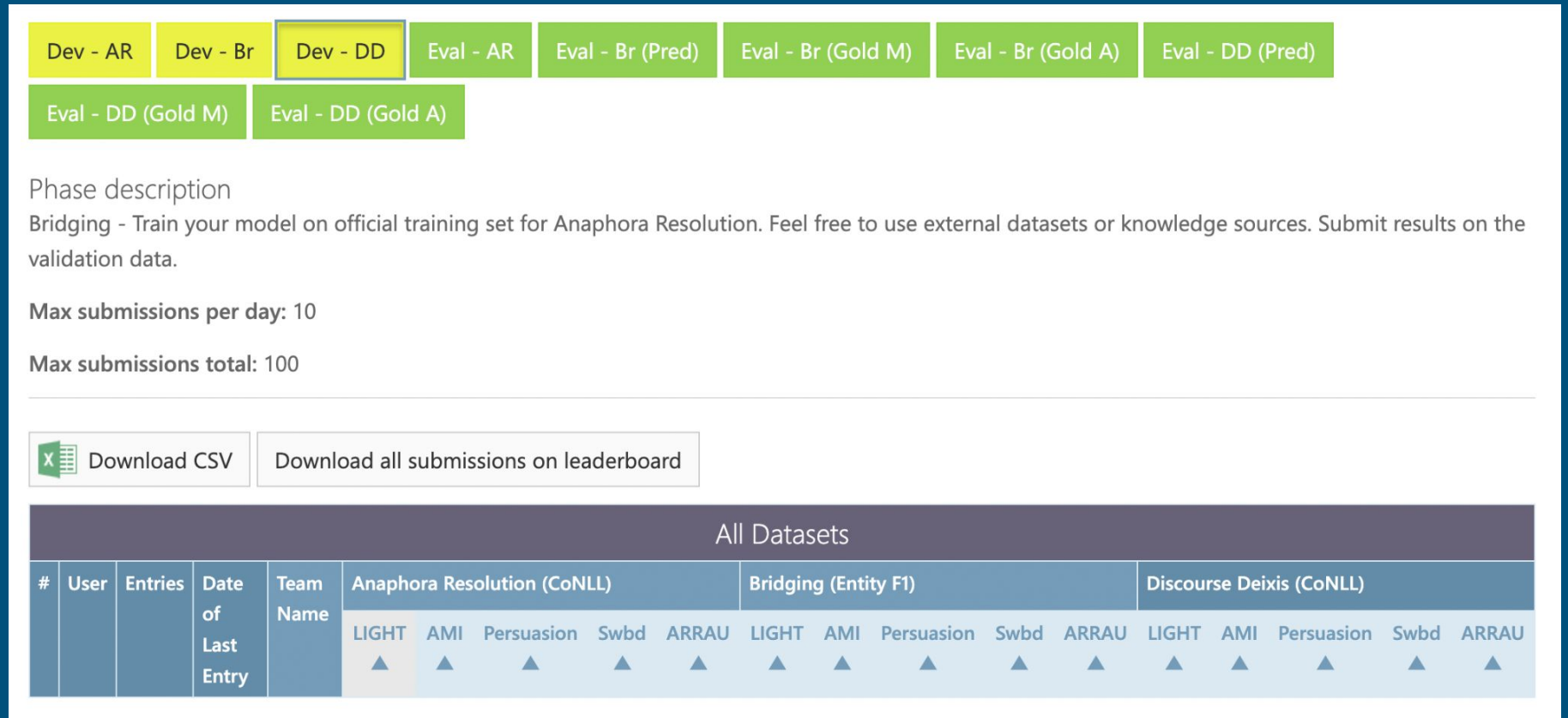
CODI-CRAC 2022: Scoring

- The Universal Anaphora Scorer [1]
  - **Identity Anaphora:** CONLL (the unweighted average of MUC, B3, CEAF)
    - `python ua-scorer.py key system`
  - **Discourse Deixis:** Same as above
    - `python ua-scorer.py key system evaluate_discourse_deixis`
  - **Bridging:** Entity F1
    - `python ua-scorer.py key system keep_bridging`

[1] https://github.com/juntaoy/universal-anaphora-scorer/

CODI-CRAC 2022: Submission Details

- Separate sub-tasks for each anaphoric relationship
  - new gold anaphora phrase in addition to gold/predicted mention phases



CODI-CRAC 2022: Baselines

- Baseline Systems [2]
  - **Identity Anaphora** (Xu and Choi, 2020)
  - **Bridging** (Yu and Poesio 2020)
  - **Discourse Deixis** (Previous utterance [‘this’, ‘that’])

[2] https://github.com/juntaoy/codi-crac2022_scripts

CODI-CRAC 2022: Participating Systems

- 54 participants registered on CodaLab
- 3 teams submitted results for Task 1 (Identity Anaphora)
- 2 teams for Task 2 (Bridging)
- 2 teams for Task 3 (Discourse Deixis)

CODI-CRAC 2022: Participating Systems

- UTD_NLP
  - Identity: Mention detection -> Coreference resolution -> **new non-referring filtering**
    - Multitask learning between MD and CR using Xu and Choi (2020)
    - Output singleton clusters and enforce dialogue-specific constraints
  - Bridging: Yu and Poesio (2020) + SpanBert
    - Dialogue-specific features
    - Exploring four different pre-training and fine-tuning strategy.
  - Discourse Deixis: anaphora detection -> antecedent resolution
    - Use heuristic and a binary classifier to deliver the anaphors
    - Resolve candidate anaphors to 10 previous utterances

CODI-CRAC 2022: Participating Systems

- KU_NLP
  - Identity: Mention Detection -> Coreference Resolution
    - Binary classifier to classify the candidate spans into mention and non-mention
    - mention pair model to find the antecedents

CODI-CRAC 2022: Participating Systems

- DFKI-INRIA
  - Identity: workspace coreference system + Xu and Choi (2020)
    - using Workspace Coreference System from last year to supply singletons and Xu and Choi (2020) for non-singleton clusters
  - Bridging: only Gold A phase
    - simplified Joshi et al. (2019) system with mention pruning and coarse-to-fine steps removed
  - Discourse Deixis: multi-task learning + Xu and Choi (2020)
    - Uses heuristics to find the candidate anaphors.
    - Then resolve the antecedents and finally uses an anaphora type classifier to filter out the identity, non-referring anaphors.

CODI-CRAC 2022: Results

- Identity Anaphora
  - 55 runs submitted to leaderboard
  - Significant improvements over baseline
  - Rank 1: UTD_NLP
    - ~15 points better than Rank 2 on SWBD
  - Toughest sub-corpus: AMI
    - 10-20 CoNLL Avg. F1 points below others

Team	LIGHT	AMI	PERS.	SWBD.	Avg.
			Eval AR		
UTD_NLP	**82.23**	**62.90**	**79.20**	**75.81**	**75.04**
DFKI-INRIA	72.06	51.41	69.87	60.61	63.49
KU_NLP	68.27	48.87	69.06	60.99	61.80
Baseline	54.23	34.14	53.16	49.30	47.71

Table 3: Performance on Task 1 (Evaluation Phase) – Identity Anaphora (CoNLL Avg. F1)

CODI-CRAC 2022: Results

- Bridging Anaphora
  - 102 runs submitted to leaderboard
  - Significant improvements over baseline
  - Rank 1: UTD_NLP
  - Higher score for Persuasion and no significant difference for other genres.
  - Gold A >> Gold M ≈ Pred
    - Performance on **Gold A** much better than **Gold M** and **Pred** setting

Team	LIGHT	AMI	PERS.	SWBD.	Avg.
Eval Br (Gold A)					
UTD_NLP	**46.80**	**39.35**	**56.91**	**44.40**	**46.87**
DFKI-INRIA	37.68	35.23	50.99	35.78	39.92
Baseline	29.93	22.69	37.83	30.39	30.21
Eval Br (Gold M)					
UTD_NLP	**26.77**	**19.65**	**34.59**	**22.74**	**25.94**
Baseline	4.99	8.77	11.49	7.08	8.08
Eval Br (Pred)					
UTD_NLP	**23.25**	**13.42**	**27.75**	**19.72**	**21.04**
Baseline	4.01	4.66	8.45	4.00	5.28

Table 4: Performance on Task 2 (Evaluation Phase) – Bridging Anaphora (Entity F1)

CODI-CRAC 2022: Results

- Discourse Deixis
  - 72 runs submitted to leaderboard
  - Significant improvements over baseline
  - Rank 1: UTD_NLP
    - small improvement Pred to Gold M
    - large improvement Gold M to Gold A
  - Toughest sub-corpora: Light

Team	LIGHT	AMI	PERS.	SWBD.	Avg.
Eval DD (Gold A)					
UTD_NLP	**52.40**	**72.50**	**69.61**	**72.11**	**66.66**
DFKI-INRIA	44.95	56.54	62.79	0.00	41.07
Baseline	40.07	39.89	51.43	37.72	42.28
Eval DD (Gold M)					
UTD_NLP	**38.38**	**55.12**	**54.89**	**49.83**	**49.56**
DFKI-INRIA	35.91	47.13	48.24	0.00	32.82
Baseline	18.14	22.95	30.15	21.37	23.15
Eval DD (Pred)					
UTD_NLP	**37.09**	**53.31**	**54.59**	**49.76**	**48.69**
DFKI-INRIA	36.82	50.09	47.04	0.00	33.49
Baseline	10.94	17.39	16.61	13.30	14.56

Table 5: Performance on Task 3 (Evaluation Phase) – Discourse Deixis (CoNLL Avg. F1)

CODI-CRAC 2022: Conclusion

- Focus on resolving three types of anaphoric relations in dialogues: identity, bridging, and discourse deixis
- Release revised annotation guideline prior to the shared task
- Release newly annotated and revised sub-corpora from different conversation genres
- Introducing new gold anaphor phases for bridging and discourse deixis

Timeline

- Mar 28 - Corrected training data and script release
- June 18 - Test data for Eval - AR, Eval - Br (Pred), and Eval - DD (Pred) released.
- July 13 - Submission deadline Eval - AR, Eval - Br (Pred), and Eval - DD (Pred).
- July 14 - Test data for Eval - Br (Gold M) and Eval - DD (Gold M) released.
- July 24 - Submission deadline Eval - Br (Gold M) and Eval - DD (Gold M).
- July 25 - Test data for Eval - Br (Gold A) and Eval - DD (Gold A) released.
- Aug 5 - Submission deadline Eval - Br (Gold A) and Eval - DD (Gold A).
- Aug 12 - System descriptions due.
- Sep 1 - Accept/reject notifications.
- Sep 15 - Camera-ready version due.
- Oct 17 - CODI-CRAC 2022 @ COLING

Thank You!