Anaphora Resolution beyond OntoNotes

Juntao Yu

Queen Mary University of London
University of Essex

DALI End of Project Workshop

Summary of My Research

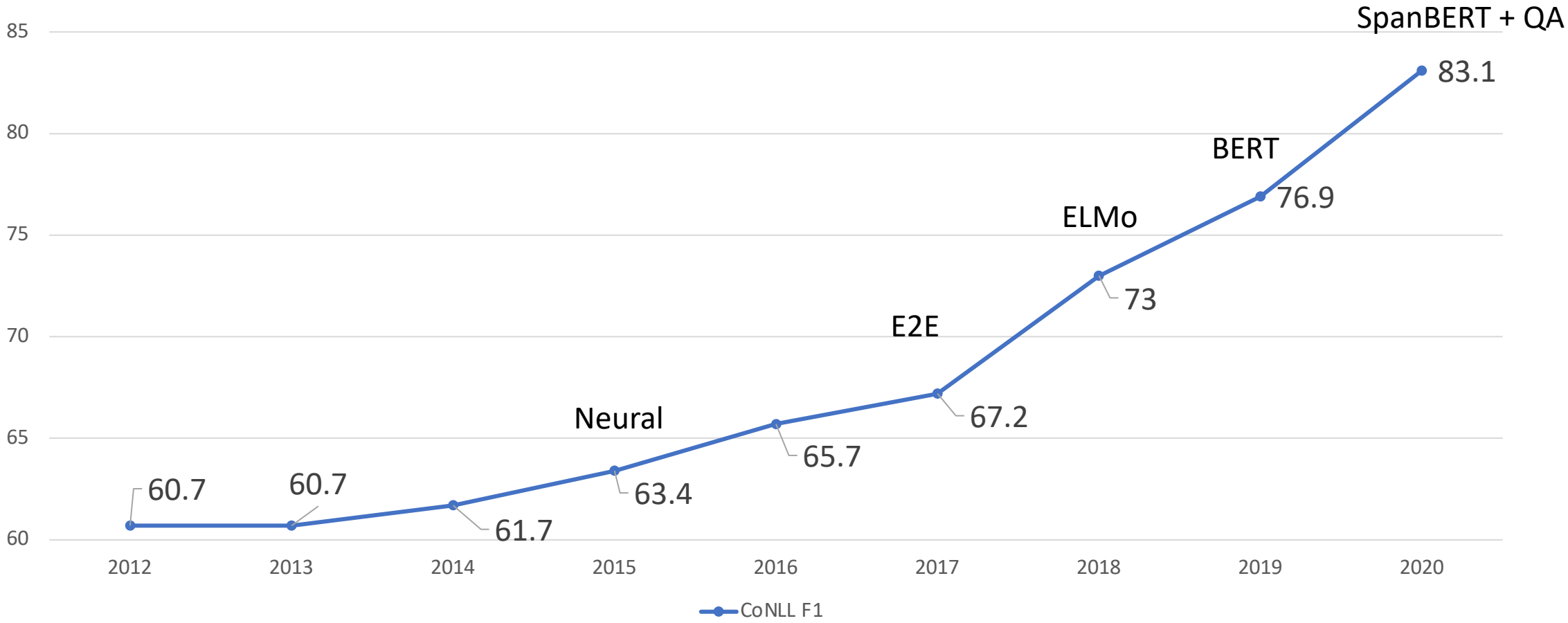
- Anaphora (Coreference) Resolution
  - Singleton Clusters/Non-referring *(Yu et al., LREC, 2020)*
  - Bridging Reference *(Yu and Poesio, COLING, 2020)*
  - Split-antecedent Plurals *(Yu et al., COLING, 2020; Yu et al., NAACL, 2021)*
  - Other topics *(Poesio et al., CRAC, 2018; Paun et al., EMNLP, 2018; Poesio et al., NAACL, 2019; Aloraini et al., CRAC, 2020)*

- Named Entity Recognition/Mention Detector
  - Neural Mention Detection *(Yu et al., LREC, 2020)*
  - NER as Dependency Parsing *(Yu et al., ACL, 2020)*
  - Annotation Game *(Madge et al., ACL, 2019; Madge et al., HCOMP, 2019)*

Schedule

- A Cluster Ranking Model for Full Anaphora Resolution
- Free the Plural: Unrestricted Split-Antecedent Anaphora Resolution
- Stay Together: A System for Single and Split-antecedent Anaphora Resolution
- Multi-task Learning Based Neural Bridging Reference Resolution

Coreference Resolution on OntoNotes



The Limitation of OntoNotes

Type	OntoNotes	ARRAU
Single-antecedent anaphors	Yes	Yes
Singletons	No	Yes
Non-referring expressions	Predicative	Yes[1]
Split-antecedent anaphors	No	Yes
Bridging reference	No	Yes
Discourse deixis	Event anaphora	Yes
Ambiguity	No	Yes

[1]Non-referring in ARRAU: predicative NPs (John is a policeman), expletives (It rains), nonreferring quantifiers (I see nobody here), idioms (He asked her for her hand) etc.

A Cluster Ranking Model for Full Anaphora Resolution

Juntao Yu, Alexandra Uma and Massimo Poesio

Queen Mary University of London

LREC 2020

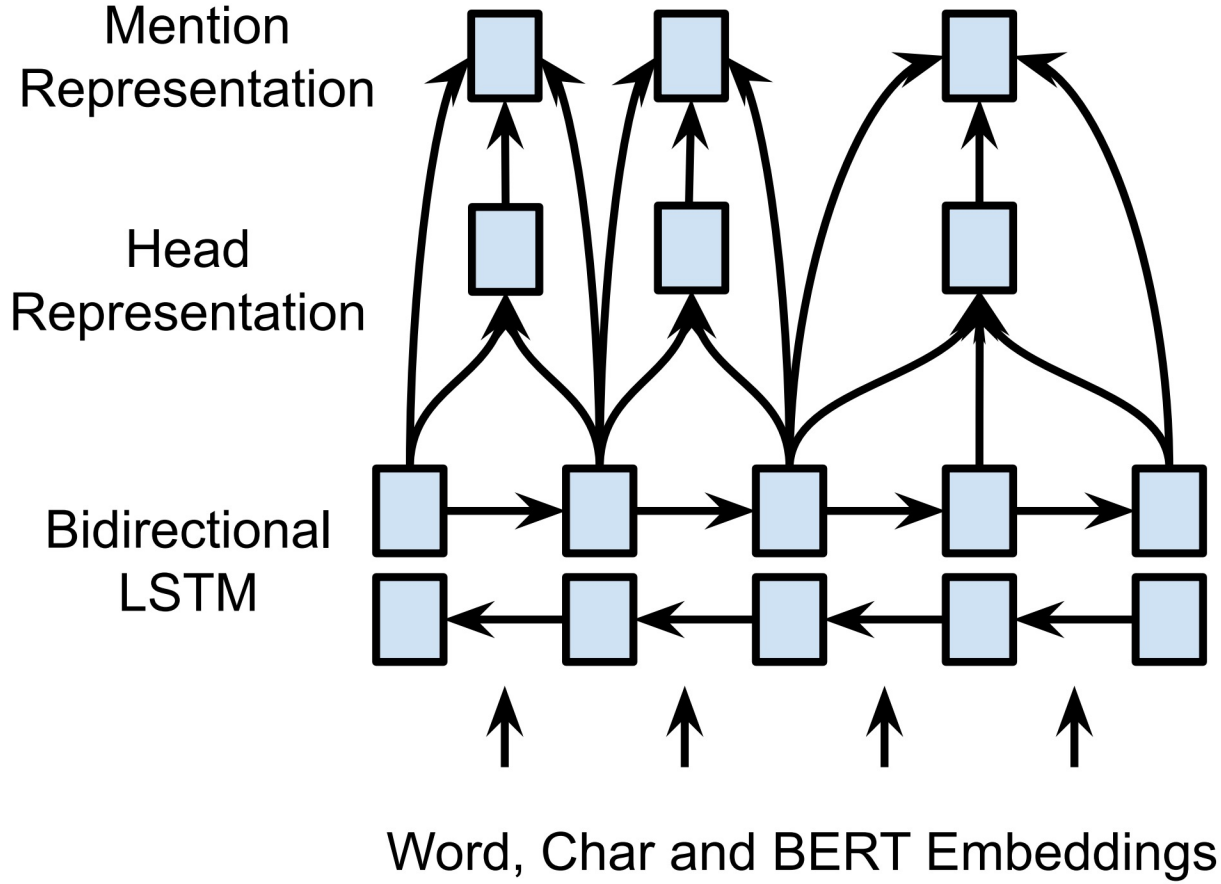
The Tasks

- Single-antecedent anaphors
- Singletons
- Non-referring expressions

Cluster vs Mention Ranking Plus

- Linguistically more appealing
- Simpler in term of the parameters (4.8M vs 9.6M)
- More flexible mention handling, e.g., prefiltering non-referring
- Explore rich cluster level features
- Support advanced search algorithms, e.g., beam search

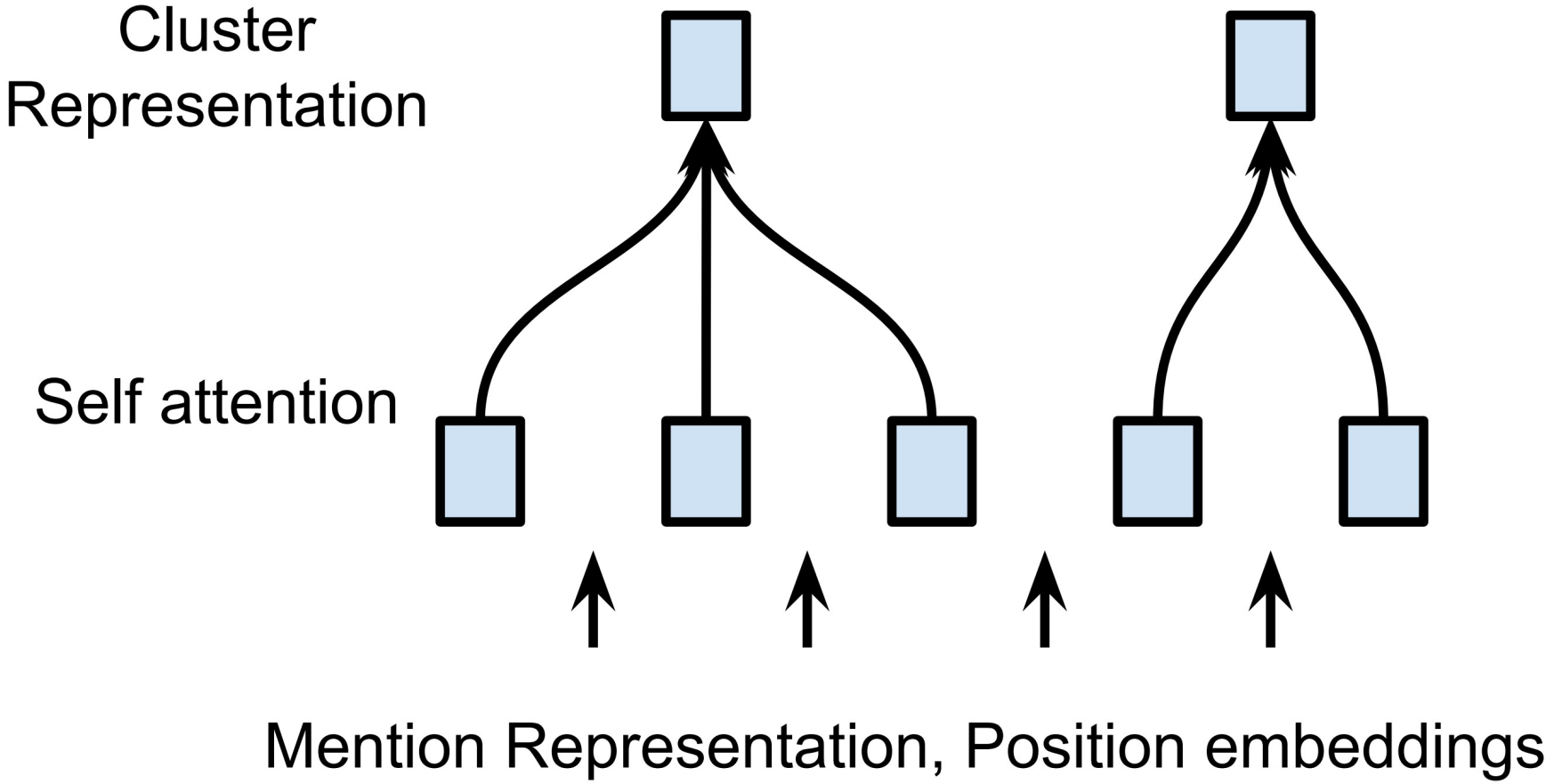
Mention Representation



Cluster Ranking

- 0.4 mention/token
- Mentions are attached directly to the clusters progressively

Cluster Representation



Cluster History

- Without cluster history:
  - Before: [Steve Jobs, he, Steve], [John], [Peter], [Tom], [Smith]
  - After: **[Steve Jobs, he, Steve, the apple founder]**, [John], [Peter], [Tom], [Smith]

- With cluster history:
  - Before: [Steve Jobs], [Steve Jobs, he], [Steve Jobs, he, Steve], [John], [Peter], [Tom], [Smith]
  - After: [Steve Jobs], [Steve Jobs, he], [Steve Jobs, he, Steve], **[Steve Jobs, he, Steve, the apple founder]**, [John], [Peter], [Tom], [Smith]

Handling Singletons and Non-referrings

$$s(i,j) = \begin{cases} s_\epsilon(i) & j = \epsilon \\ s_m(i) + s_c(j) + s_{mc}(i,j) & j \neq \epsilon \end{cases}$$

$$s_\epsilon(i) = \begin{cases} s_{no}(i) & \mathbf{NO} \\ s_{nr}(i) + s_m(i) & \mathbf{NR} \\ s_{dn}(i) + s_m(i) & \mathbf{DN} \end{cases}$$

Non-referring expressions will be filtered out once selected by the prefiltering or hybrid rules.

Results on ARRAU

	Models	MUC			B^3			$CEAF_{\phi_4}$			Avg. F1
		P	R	F1	P	R	F1	P	R	F1	
Singletons included	PREFILTERING	75.5	**79.0**	77.2	75.9	**80.7**	78.2	75.2	77.3	76.2	77.2
	HYBRID	**77.9**	78.5	**78.2**	**77.4**	80.3	**78.8**	**75.4**	78.1	**76.8**	**77.9**
	FINE NR	76.7	77.3	77.0	76.8	79.7	78.2	74.9	78.0	76.4	77.2
	Lee et al. (2013)*	72.1	58.9	64.8	77.5	77.1	77.3	64.2	**88.1**	74.3	72.1
Singletons excluded	PREFILTERING	75.5	**79.0**	77.2	67.0	**73.0**	69.9	67.1	**65.1**	66.1	71.1
	HYBRID	**77.9**	78.5	**78.2**	**69.2**	71.8	**70.4**	**69.5**	63.8	**66.5**	**71.7**
	FINE NR	76.7	77.3	77.0	68.0	70.7	69.3	66.6	64.2	65.4	70.6
	NO NR	76.7	77.0	76.8	68.7	69.7	69.2	66.1	63.8	64.9	70.3
	Lee et al. (2013)*	72.3	58.9	64.9	67.9	48.5	56.5	54.2	53.0	53.6	58.3

Table 2: The comparison between our models and the SoTA system on the CRAC test set. * indicates systems evaluated on the gold mentions.

Results on ARRAU

Models	P	R	F1
PREFILTERING	76.6	74.5	75.5
HYBRID	78.0	72.4	75.1
FINE NR	77.0	75.5	76.3

Table 3: The scores for non-referring expressions of our models on the CRAC test set.

NR types	P	R	F1
Expletive	93.8	100.0	96.8
Predicate	77.6	75.2	76.4
Quantifier	65.0	64.7	64.9
Coordination	77.5	82.0	79.7
Idiom	77.0	55.9	64.8

Table 4: The scores of our models on the fine-grained non-referring types.

Results on OntoNotes

	Models	MUC			B^3			$CEAF_{\phi_4}$			Avg. F1
		P	R	F1	P	R	F1	P	R	F1	
Context Independent Embeddings	Clark and Manning (2016a)	79.2	70.4	74.6	**69.9**	58.0	63.4	63.5	55.5	59.2	65.7
	Lee et al. (2017)	78.4	73.4	75.8	68.6	61.8	65.0	62.7	**59.0**	60.8	67.2
	Zhang et al. (2018)	**79.4**	**73.8**	**76.5**	69.0	**62.3**	**65.5**	**64.9**	58.3	**61.4**	**67.8**
Pre-trained Contextual Embeddings	Lee et al. (2018)	81.4	79.5	80.4	72.2	69.5	70.8	68.2	67.1	67.6	73.0
	Kantor and Globerson (2019)	82.6	**84.1**	**83.4**	73.3	**76.2**	**74.7**	**72.4**	**71.1**	**71.8**	**76.6**
	Our model	**82.7**	83.3	83.0	**73.8**	75.6	**74.7**	72.2	71.0	71.6	76.4
Fine-tuned on BERT	Joshi et al. (2019b)	84.7	82.4	83.5	76.5	74.0	75.3	74.1	69.8	71.9	76.9
	Joshi et al. (2019a)	**85.8**	**84.8**	**85.3**	**78.3**	**77.9**	**78.1**	**76.4**	**74.2**	**75.3**	**79.6**

Table 6: Comparison between our models and the top performing systems on the CONLL test set.

Free the Plural: Unrestricted Split-Antecedent Anaphora Resolution

Juntao Yu[1] , Nafise Sadat Moosavi[2] , Silviu Paun[1] , Massimo Poesio[1]

[1]Queen Mary University of London
[2]UKP Lab, Technische Universitat Darmstadt

COLING 2020

Single/Split Antecedent Plurals

- Single-antecedent
  - The *Joneses*$_i$ went to the park. <u>They</u>$_i$ had a good time.
  - *John and Mary*$_i$ went to the park. <u>They</u>$_i$ had a good time.

- Split-antecedent (Eschenbach et al., 1989)
  - *John*$_i$ met *Mary*$_j$ in the park. <u>They</u>$_{i,j}$ had a good chat .
  - John likes *green*$_i$, Mary likes *blue*$_j$, but Tom likes <u>both colours</u>$_{i,j}$.

The work on split-antecedent plurals are limited

Challenge: Under-Resourced

Corpus	Anaphors Type	Data Quality	Num of docs	Num of Anaphors
ARRAU TRAIN/DEV/TEST	Split-antecedent anaphors	Gold	211 / 30 / 60	507 / 80 / 110

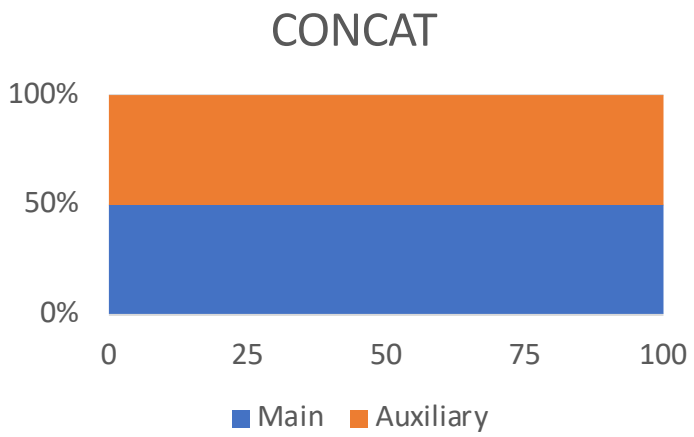
Challenge: Under-Resourced

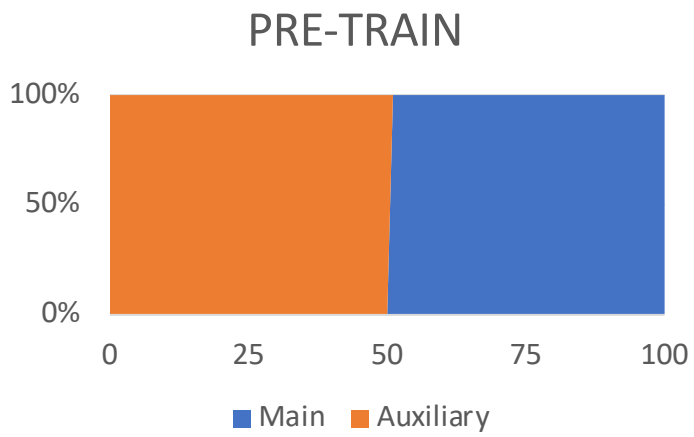
Corpus	Anaphors Type	Data Quality	Num of docs	Num of Anaphors
ARRAU TRAIN/DEV/TEST	Split-antecedent anaphors	Gold	211 / 30 / 60	507 / 80 / 110
PD-SILVER	Split-antecedent anaphors	Silver	165	507
PD-CROWD	Split-antecedent anaphors	Noisy	467	6262
ELEMENT-OF	Bridging anaphors	Gold	213	1059
SINGLE-COREF	Single-antecedent anaphors	Gold	462	30372

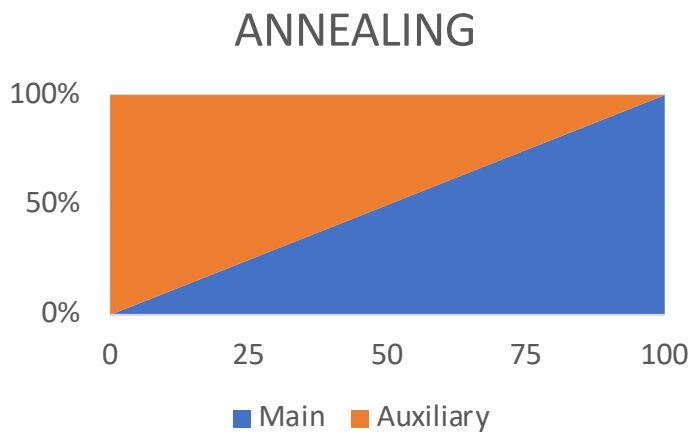
Table 1: Statistics about the corpora used in our experiments.

Training Strategies

- Concatenation (CONCAT)
- Pre-training (PRE-TRAIN)
- Corpus Annealing (ANNEALING)







We train our system on a simplified version of Lee et al., (2018); Kantor and Globerson (2019)

Training Strategies Selection

Training Strategy	BASELINE	PD-SILVER	PD-CROWD	ELEMENT-OF	SINGLE-COREF
CONCAT	58.2	**59.8**	61.2	59.2	67.6
PRE-TRAIN	58.2	59.0	**62.9**	**64.3**	66.5
ANNEALING	58.2	59.0	62.6	61.1	**69.5**

Table 2: Training strategy selection on the development set with lenient F1 scores.

Results

	Lenient			Strict
	R	**P**	**F1**	**Accuracy**
RECENT-2	19.6	21.8	20.6	3.6
RECENT-3	31.8	23.6	27.1	0.9
RECENT-4	40.4	22.6	28.9	0.0
RECENT-5	45.7	20.4	28.2	0.0
RANDOM	24.9	11.4	15.7	0.0
NEURAL BASELINE	60.8	56.4	58.6	22.7
PD-SILVER	61.6	61.9	61.8	30.9
PD-CROWD	68.2	63.5	65.7	31.8
ELEMENT-OF	64.5	65.0	64.8	34.5
SINGLE-COREF	68.6	**70.6**	69.6	42.7
PD-CROWD + SINGLE-COREF	68.2	69.6	68.9	40.9
ELEMENT-OF + SINGLE-COREF	69.4	67.5	68.4	39.1
PD-CROWD + ELEMENT-OF + SINGLE-COREF	**72.2**	67.8	**70.0**	**43.6**

Table 3: Comparing our models with the baselines on the test set.

Stay Together: A System for Single and Split-antecedent Anaphora Resolution

Juntao Yu[1] , Nafise Sadat Moosavi[2] , Silviu Paun[1] , Massimo Poesio[1]

[1]Queen Mary University of London
[2]UKP Lab, Technische Universitat Darmstadt

NAACL 2021

Handling Split-antecedents

$$s_p(i,j) = \begin{cases} 0 & j = \epsilon \\ s_m(i) + s_c(j) + s_{pmc}(i,j) & j \in C_{i-1} \end{cases}$$

$$p_p(i,j) = \frac{1}{1 + e^{-s_p(i,j)}}$$

Training

- Pre-Trained

$$loss_p = log \prod_{j=1}^{N} \sum_{\hat{c} \in \text{GOLD}_p(j)} s_p(\hat{c}, j)$$

- Joint

$$loss_j = (1 - \beta) loss_s + \beta loss_p$$

Results on ARRAU RST (Separate Evaluation)

	CoNLL	**Non-referring**			**Anaphora Rec$_{split}$**			**Lenient$_{split}$**			**Strict$_{split}$**		
	F1	**R**	**P**	**F1**	**R**	**P**	**F1**	**R**	**P**	**F1**	**R**	**P**	**F1**
Recent-2	-	-	-	-	31.7	42.2	36.2	10.3	15.6	12.4	5.0	6.7	5.7
Recent-3	-	-	-	-	31.7	42.2	36.2	16.9	17.0	17.0	0.0	0.0	0.0
Recent-4	-	-	-	-	30.0	40.9	34.6	18.4	14.2	16.0	0.0	0.0	0.0
Recent-5	-	-	-	-	28.3	39.5	33.0	16.9	10.7	13.1	0.0	0.0	0.0
Random	-	-	-	-	31.7	42.2	36.2	5.9	3.7	4.5	0.0	0.0	0.0
JOINT	77.1	**72.6**	77.2	74.8	**50.0**	51.7	50.9	**39.0**	35.3	**37.1**	15.0	15.5	15.3
PRE-TRAINED	**77.9**	72.4	**78.0**	**75.1**	45.0	**71.1**	**55.1**	30.2	**46.1**	36.4	**16.7**	**26.3**	**20.4**

Table 3: Separate evaluation of our systems on the test set. X$_{split}$ are the scores for the split-antecedent anaphors.

Joint Evaluation use the Extended LEA Metric

$$\frac{\sum_{e \in E} \text{importance}(e) * \text{resolution-score}(e)}{\sum_{e \in E} \text{importance}(e)}$$

$$RS(e) = \frac{1}{|\mathbb{L}(e)|} \sum_{l \in \mathbb{L}(e)} \mathbb{B}(l, K)$$

$$\mathbb{B}(l, K) =$$

$$\begin{cases} \frac{|s_i| * |s_j|}{|m_i| * |m_j|} & \{\exists_{k \in K, s_i \in \hat{\mathbb{P}}(m_i), s_j \in \hat{\mathbb{P}}(m_j)} | l_{s_i, s_j} \in \mathbb{L}(k)\} \\ \frac{|m_i| * |m_j|}{|m_k| * |m_p|} & \{\exists_{k \in K, m_i \in \hat{\mathbb{P}}(m_k), m_j \in \hat{\mathbb{P}}(m_p)} | l_{m_k, m_p} \in \mathbb{L}(k)\} \\ 0 & \text{otherwise} \end{cases}$$

$$\text{importance}(e) = \frac{\text{importance-factor}(e) * |e|}{\sum_{e_i} \text{importance-factor}(e_i) * |e_i|}$$

$$\text{importance-factor}(e) = \begin{cases} \text{Imp}_{split} & \text{If } e \text{ is a plural entity} \\ 1 & \text{If } e \text{ is singular} \end{cases}$$

Results on ARRAU RST (Joint Evaluation)

	$\textbf{Imp}_{split}$ **= 1**			$\textbf{Imp}_{split}$ **= 10**		
	R	**P**	**F1**	**R**	**P**	**F1**
Recent-2	70.5	66.9	68.7	61.5	61.3	61.4
Recent-3	70.5	66.9	68.7	61.6	61.1	61.4
Recent-4	70.6	66.9	68.7	61.8	61.1	61.5
Recent-5	70.5	66.9	68.7	61.5	61.2	61.3
Random	70.4	66.7	68.5	60.9	60.0	60.4
Our model	**70.8**	**67.2**	**69.0**	**63.8**	**64.4**	**64.1**

Table 4: LEA evaluation on both single- and split-antecedent anaphors. Imp_{split} indicates the split-antecedent importance.

Multitask Learning-Based Neural Bridging Reference Resolution

Juntao Yu and Massimo Poesio
Queen Mary University of London

COLING 2020

Bridging References

- Bridging references (Hawkins, 1974; Clark, 1975; Prince, 1981, 1990) are references related to a discourse model entity by an associative relation whose recognition requires some ’bridging inference’

  - [The Bakersfield Supermarket] went bankrupt last May. [The business] closed when [[its] old owner] was murdered by [robbers]. [The murder] saddened [the customers].

  - Coreference: [The Bakersfield Supermarket] [The business] [its]
  - Bridging: [the customers] → [The Bakersfield Supermarket]

Example adapted from Hou et al., (2018)

Challenge No 1: Different Annotation Schemes

- Bridging based on relational nouns (called `Referential' by e.g., Rosiger et al, 2018)
  - The bridging nominal has an implicit anaphoric argument
  - John walked towards [the house]. The door was open.

- More general bridging (called `Lexical' bridging by Rosiger et al, 2018)
  - The bridging nominal could also be interpreted autonomously
  - I went to [Spain] last year. I particularly liked Madrid.

Challenge No 2: Sparse Data

Corpus	Genre	Bridging Type	Mention Type	Number of Bridging
ARRAU RST	WSJ news	lexical, referential bridging	Gold	3303
ARRAU TRAINS	Dialogues	lexical, referential bridging	Gold	558
ARRAU PEAR	Narrative	lexical, referential bridging	Gold	303
ISNOTES	WSJ news	referential bridging	Gold	663
BASHI	WSJ news	referential bridging	Predicted	459
SCICORP	Scientific texts	referential bridging	Predicted	1366

Table 1: The statistics of bridging corpora used in our experiments.

The Network Architectures

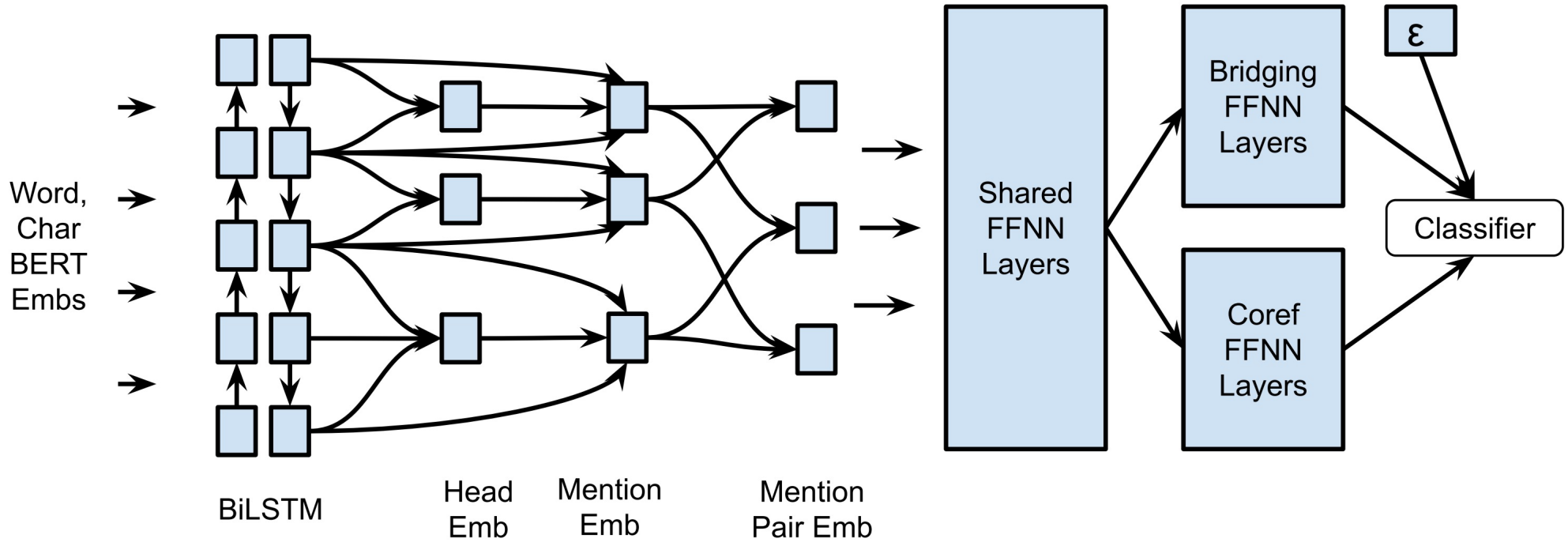


Figure 1: The proposed multi-task architecture.

Parameter Tuning

System	Shared Network	RST	ISNOTES
bridging only		47.4	33.8
multi-task	embeddings, LSTM	50.9	38.7
	+ 1 FFNN Layer	**54.7**	**43.7**
	+ 2 FFNN Layer	51.7	40.1

(a) antecedent selection

	RST						ISNOTES					
	anaphor rec.			full bridging res.			anaphor rec.			full bridging res.		
System	P	R	F1	P	R	F1	P	R	F1	P	R	F1
bridging only	50.0	12.5	20.0	34.5	8.6	13.8	63.9	16.2	25.8	33.3	8.5	13.5
multi-task	47.3	19.0	27.1	35.5	14.2	20.3	59.3	22.5	32.7	31.5	12.0	17.4
+ undersampling	31.5	36.2	**33.7**	20.6	23.7	**22.0**	45.6	47.2	**46.4**	19.1	19.7	**19.4**

(b) full bridging resolution

Table 3: Parameter tuning on the dev set of ARRAU RST and ISNOTES.

Results on Antecedent Selection

	RST	TRAINS	PEAR	ISNOTES[7]	BASHI	SCICORP
Hou et al. (2013)	-	-	-	41.3	-	-
Hou (2018a)	32.4	-	-	46.5	27.4	-
Roesiger (2018)	39.8	48.9	28.2	-	-	
Hou (2020)	34.6	-	-	**50.1**	-	-
Our model	**49.3**	**50.9**	**61.2**	43.7	**36.0**	**33.4**

Table 5: Comparing our model with the SoTA for antecedent selection.

Results on Full Bridging

Corpus	Gold Coreference Anaphors Setting	Models	anaphor rec.			full bridging res.		
			P	R	F1	P	R	F1
RST	Keep	Our model	31.8	29.8	30.8	20.2	18.9	19.5
	Remove	Roesiger (2018)	29.2	32.5	30.7	18.5	20.6	19.5
		Our model	37.6	35.9	**36.7**	24.6	23.5	**24.0**
TRAINS	Keep	Our model	49.4	36.0	41.6	33.7	24.6	28.4
	Remove	Roesiger (2018)	39.3	21.8	24.2	27.1	21.8	24.2
		Our model	62.2	40.4	**48.9**	39.2	25.4	**30.9**
PEAR	Keep	Our model	74.7	48.8	59.0	68.4	44.6	54.0
	Remove	Roesiger (2018)	75.0	16.0	26.4	57.1	12.2	20.1
		Our model	81.1	49.6	**61.5**	74.3	45.5	**56.4**
ISNOTES	Keep	Hou et al. (2014) [8]	65.9	14.1	23.2	57.7	10.1	17.2
		Roesiger et al. (2018)	45.9	18.3	26.2	32.0	12.8	18.3
		Our model	53.9	33.6	**41.4**	31.6	19.8	**24.3**
	Remove	Roesiger et al. (2018)	71.6	18.3	29.2	50.0	12.8	20.4
		Our model	58.3	35.1	**43.8**	33.5	20.2	**25.2**
BASHI	Keep	Our model	34.4	34.2	34.3	17.7	17.5	17.6
	Remove	Roesiger et al. (2018)	49.4	20.2	28.7	24.3	10.0	14.1
		Our model	35.3	34.9	**35.1**	18.2	18.0	**18.1**
SCICORP	Keep	Our model	45.0	35.7	39.8	21.5	17.1	19.0
	Remove	Roesiger et al. (2018)	17.7	0.9	8.1	3.2	0.9	1.5
		Our model	52.9	41.2	**46.3**	25.0	19.4	**21.9**

Table 6: Comparing our model with the SoTA for full bridging resolution.

The CODI-CRAC 2021 Shared Task on Anaphora, Bridging, and Discourse Deixis in Dialogue

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

CODI-CRAC@EMNLP 2021

UA Scorer https://github.com/juntaoy/universal-anaphora-scorer

Universal Anaphora in CONLL-U-Plus Format

```
# global.columns = ID FORM LEMMA UPOS XPOS FEATS HEAD DEPREL DEPS MISC IDENTITY BRIDGING DISCOURSE_DEIXIS REFERENCE NOM_SEM
# newdoc id = Trains_93/D93_9_1
# turn_id = D93_9_1-t1
# speaker = s
# sent_id = D93_9_1-1
# text = s : hello can I help you
1     s       _  _  _  _  _  _  _  _  _  _  _  _
2     :       _  _  _  _  _  _  _  _  _  _  _  _
3     hello   _  _  _  _  _  _  _  _  _  _  _  _
4     can     _  _  _  _  _  _  _  _  _  _  _  _
5     I       _  _  _  _  _  _  _  _  _  _  _  _
6     help    _  _  _  _  _  _  _  _  _  _  _  _
7     you     _  _  _  _  _  _  _  _  _  _  _  _

# turn_id = D93_9_1-t2
# speaker = u
# sent_id = D93_9_1-2
# text = u : okay um
8     u       _  _  _  _  _  _  _  _  _  _  (EntityID=1-DD|MarkableID=dd_markable_535|Min=8,23|SemType=dn  _
9     :       _  _  _  _  _  _  _  _  _  _  _                                                           _
10    okay    _  _  _  _  _  _  _  _  _  _  _                                                           _
11    um      _  _  _  _  _  _  _  _  _  _  _                                                           _

# sent_id = D93_9_1-3
# text = I want to know how long alright how long does it take
12    I             _  _  _  _  _  _  _  _  _                                                                  _  _  _
13    want          _  _  _  _  _  _  _  _  _                                                                  _  _  _
14    to            _  _  _  _  _  _  _  _  _                                                                  _  _  _
15    know          _  _  _  _  _  _  _  _  _                                                                  _  _  _
16    how           _  _  _  _  _  _  _  _  (EntityID=1-Pseudo|MarkableID=markable_469|Min=16,17|SemType=quantifier  _  _  _  (MarkableID=markable_469|Entity_Type=unknown|Genericity=no-generic
17    long          _  _  _  _  _  _  _  _  )                                                                  _  _  _  )
18    alright       _  _  _  _  _  _  _  _  _                                                                  _  _  _
19    how           _  _  _  _  _  _  _  _  (EntityID=2-Pseudo|MarkableID=markable_470|Min=19,20|SemType=quantifier  _  _  _  (MarkableID=markable_470|Entity_Type=unknown|Genericity=no-generic
20    long          _  _  _  _  _  _  _  _  )                                                                  _  _  _  )
21    does          _  _  _  _  _  _  _  _  _                                                                  _  _  _
22    it            _  _  _  _  _  _  _  _  (EntityID=3-Pseudo|MarkableID=markable_6|Min=22|SemType=expletive)       _  _  _  (MarkableID=markable_6|Entity_Type=unknown|Genericity=no-generic)
23    take          _  _  _  _  _  _  _  _  _                                                                  _  )  _
```

https://github.com/UniversalAnaphora/UniversalAnaphora/blob/main/UA_CONLL_U_Plus_proposal_v1.0.md

Evaluation Modes

- Evaluating coreference relations only
- Evaluating coreference relations (include split-antecedents) and singletons
- Evaluating all markables (exclude discourse-deixis)
- Evaluating discourse-deixis
- Evaluating only split-antecedent alignment
- Minimum Span Evaluation

Code

- The code and pre-trained models are available from our GitHub pages:
- https://github.com/juntaoy
- https://github.com/dali-ambiguity

Thank You!