Event Argument Evaluation

Marjorie Freedman (ISI)
Ryan Gabbard (ISI)
Jay DeYoung (BBN)

Outline

- Overview of EAL Task
- Participants & Approaches
- 2017 Results

Event Argument Task

Event Argument Task

In a document

- Identify what events occurred along with their type
- Identify key arguments (e.g. participants, dates, locations) and associate them with the correct events
- Provide arguments realis status (ACTUAL, OTHER, GENERIC)
- Group arguments into event hoppers

A separatist group called the Kurdistan Freedom Falcons (TAK) claimed responsibility for an explosion late on Monday which wounded six people, one of them seriously, in an Istanbul supermarket. Istanbul governor Muammer Guler told Anatolia news agency the explosion in the Bahcelievler district of Turkey's largest city injured six people. The agency said 15 other people had been hurt. "We consider the explosion that took place tonight in an Istanbul supermarket to be a response to the barbaric policies against the Kurdish people

Event2: Conflict. Attack	Role	Fillers
	ATTACKER	TAK
	TARGET	Six people 15 other people
	PLACE	the Bahcelievler district Istanbul An Istanbul supermarket
	DATE	Monday (2006-02-13)

Event1: Life.Injure	Role	Fillers
	Agent	TAK
	Victims	Six people 15 other people
	PLACE	the Bahcelievler district Istanbul An Istanbul supermarket
	DATE	Monday (2006-02-13)

2017 Event Ontology

EAL Event Label (Type.Subtype)	Role	Allowable ARG Entity/Filler Type
Conflict.Attack	Attacker	PER, ORG, GPE
	Instrument	WEA, VEH, COM
	Target	PER, GPE, ORG, VEH, FAC, WEA, COM
Conflict.Demonstrate	Entity	PER, ORG
Contact.Broadcast	Audience	PER, ORG, GPE
	Entity	PER, ORG, GPE
Contact.Contact	Entity	PER, ORG, GPE
Contact.Correspondence	Entity	PER, ORG, GPE
Contact.Meet	Entity	PER, ORG, GPE
Justice.Arrest-Jail	Agent	PER, ORG, GPE
	Crime	Crime
	Person	PER
Life.Die	Agent	PER, ORG, GPE
	Instrument	WEA, VEH, COM
	Victim	PER
Life.Injure	Agent	PER, ORG, GPE
	Instrument	WEA, VEH, COM
	Victim	PER
Manufacture.Artifact	Agent	PER, ORG, GPE
	Artifact	VEH, WEA, FAC, COM
	Instrument	WEA, VEH, COM

EAL Event Label (Type.Subtype)	Role	Allowable ARG Entity/Filler Type
Movement.Transport-Artifact	Agent	PER, ORG, GPE
	Artifact	WEA, VEH, FAC, COM
	Destination	GPE, LOC, FAC
	Instrument	VEH, WEA
	Origin	GPE, LOC, FAC
Movement.Transport-Person	Agent	PER, ORG, GPE
	Artifact	PER
Personnel.Elect	Agent	PER, ORG, GPE
	Person	PER
	Position	Title
Personnel.End-Position	Entity	ORG, GPE
	Person	PER
	Position	Title
Personnel.Start-Position	Entity	ORG, GPE
	Person	PER
	Position	Title
Transaction.Transaction	Beneficiary	PER, ORG, GPE
	Giver	PER, ORG, GPE
	Recipient	PER, ORG, GPE
Transaction.Transfer-Money	Beneficiary	PER, ORG, GPE
	Giver	PER, ORG, GPE
	Money	MONEY
	Recipient	PER, ORG, GPE
Transaction.Transfer-Ownership	Beneficiary	PER, ORG, GPE
	Giver	PER, ORG, GPE
	Recipient	PER, ORG, GPE
	Thing	VEH, WEA, FAC, ORG,COM

2017 Event Ontology

EAL Event Label (Type.Subtype)	Role	Allowable ARG Entity/Filler Type
Conflict.Attack	Attacker	PER, ORG, GPE
	Instrument	WEA, VEH, COM
	Target	PER, GPE, ORG, VEH, FAC, WEA, COM
Conflict.Demonstrate	Entity	PER, ORG
Contact.Broadcast	Audience	PER, ORG, GPE
	Entity	PER, ORG, GPE
Contact.Contact	Entity	PER, ORG, GPE
Contact.Correspo		
Contact.Meet		
Justice.Arrest-Jail		
Life.Die		
Life.Injure		
Manufacture.Arti		

EAL Event Label (Type.Subtype)	Role	Allowable ARG Entity/Filler Type
Movement.Transport-Artifact	Agent	PER, ORG, GPE
	Artifact	WEA, VEH, FAC, COM
	Destination	GPE, LOC, FAC
	Instrument	VEH, WEA
	Origin	GPE, LOC, FAC
Movement.Transport-Person	Agent	PER, ORG, GPE
	Artifact	PER
Personnel.Elect	Agent	PER, ORG, GPE
	Person	PER

Event types and subtypes the same as:

- Event nugget evaluation
- 2016 event argument evaluation

2-5 potential event-specific argument roles per event + DATE & LOCATION for all events

- Not all arguments need to be known
- Arguments can be
  - Dates, EDL entity types, string fillers (e.g. *crime*)
  - Named OR underspecified (e.g. *the unnamed suspect*)

What is Required to Fill an Event Frame

1. Finding events, arguments, and their roles (2014 task)
   A. Recognize the presence of the event → *overlap with the event nugget task but no requirement that the exact phrase is found; instead allow sentence length justifications*
   B. Find a mention (base filler) where the participation in the event (along with the role) is clear → *similar to mention level argument extraction as in event detection in ACE*
   C. Link the base filler to a canonical argument string → *use within document coreference and temporal resolution; similar to ColdStart requirement that slot-fills reference a named entity (and not a local mention)*
   D. Assign a realis label to assertion about the event and argument → *overlap with the event nugget task, but also incorporate understanding of the argument itself (e.g. failed participation)*
2. Link the argument assertions such that arguments that correspond to the same "real world" event are grouped together (Added in 2015)

Chronology of EAL Task

	Information Target	Scoring Method	Submission	Lang
2014	Table of arguments	Assessment	EAL file	En
2015	1. Table of arg. + role 2. Arg. + role grouped into frames	Assessment	EAL file	En Ch
2016	1. Table of arg. + role 2. Arg. + role grouped into frames 3. Corpus-level frame co-reference	Gold Standard for 1 & 2 Assessment for 3	EAL file	En Ch Sp
2017	1. Table of arg. + role 2. Arg. + role grouped into frames	Gold Standard	EAL file *or* ColdStart++ KB	En Ch Sp

2017 Reference Data (1)

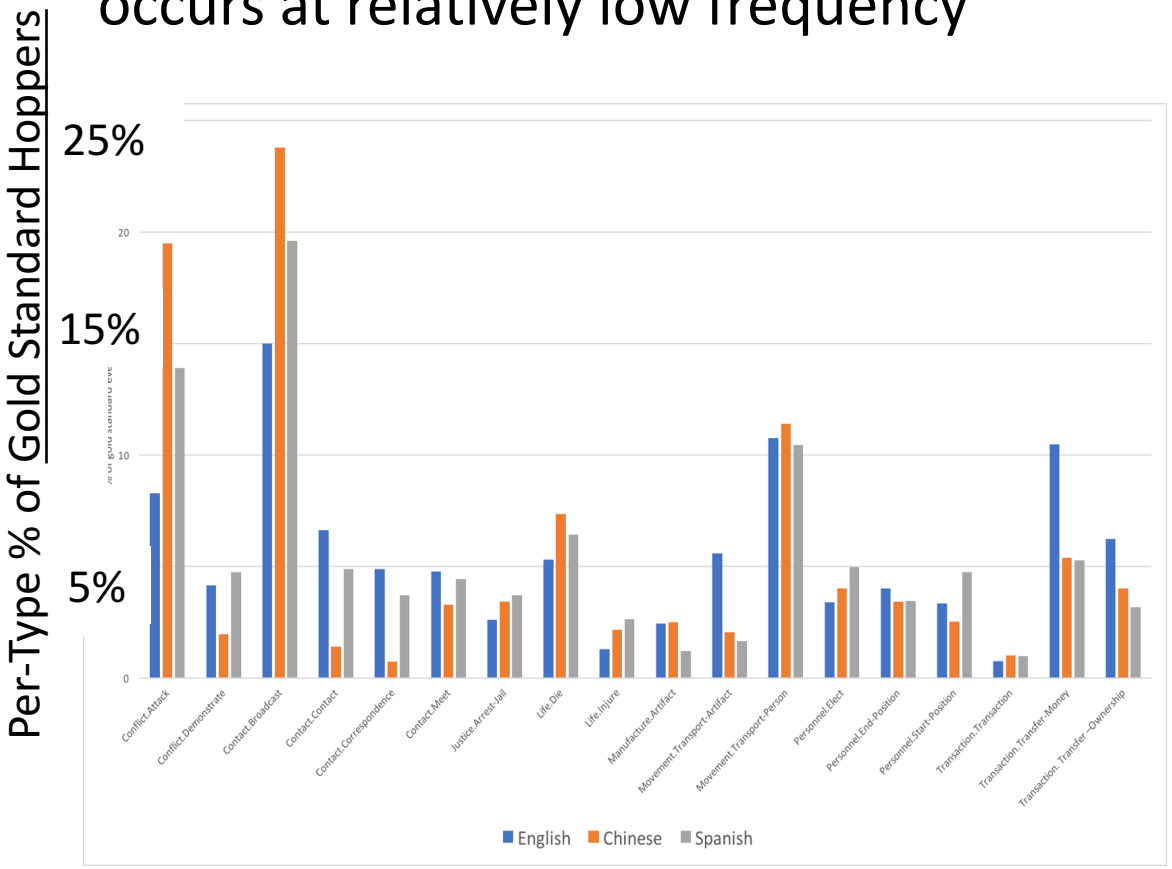
- Relied on the shared Rich ERE document set
  - ~80 documents per language
- Languages differ in
  - Total number of event hoppers
  - Average number of arguments per hopper

	# Hop.	# Arg.	Avg. Arg. per Hopper
English	2,952	7,845	2.7
Chinese	2,487	5,518	2.2
Spanish	2,049	5,917	2.9

Number of Hoppers and Arguments in the Gold Standard Reference

2017 Reference Data (2)

- With a few exceptions, relatively even distribution over 30 event types
  - Broadcast and Attack events are particularly frequent in Chinese documents
- Overall, many event types each of which occurs at relatively low frequency



	Ev. Subtype	#	%
English	Transport-Person	1,264	16%
	Broadcast	832	11%
	Transfer-Money	770	10%
	Arrest-Jail	215	3%
	Injure	88	1%
	Trans.Transaction	88	1%
Chinese	Broadcast	1,047	19%
	Attack	958	17%
	Transport-Person	727	13%
	Cont.Contact	82	1%
	Transaction	57	1%
	Correspondence	40	1%
Spanish	Transport-Person	956	16%
	Attack	780	13%
	Broadcast	700	12%
	Artifact	123	2%
	Injure	109	2%
	Trans.Transaction	91	2%

Most & Least Frequent Event Types of Event Argument Assertions

Participants & Approaches

Participants & Type of Submission

Site	EN	CH	SP	Sub
A2KD_Adept	X	X		CS++
ISCAS_Sogou		X		CS++
SAFT_ISI	X	X	X	CS++
Tinkerbell	X	X	X	CS++
BBN	X	X	X	EAL
BUPT_PRIS	X			EAL
CMU CS	X	X	X	EAL

Cold Start++	EAL
July evaluation window	Sept evaluation window
Process full ColdStart corpus (30K docs per language)	Process shared subset (~80 docs per language)
EAL valid files extracted from KB by a NIST script	EAL files submitted directly by participant
Performance measured in • Cold Start queries • EDL • EAL	Only EAL performance is measured

Approaches to Argument Assertions

… She will attend the conference. Next week's meeting …. →
(Contact.Meet, Participant, she=Marjorie Freedman, Other)
(Contact.Meet, Date, next week=W48-207, Other)

- Finding arguments: typically, pipeline approach to (1) detect triggers and (2) find arguments, exceptions:
  - **BBN**: joint inference over triggers and arguments by using a low threshold to over predict triggers
  - **BUPT_PRIS**: joint-attention based model
- Resolving arguments (e.g. co-reference, date resolution)
  - Ignored by some systems → hurts system performance
  - Core NLP coreference used by many
- Labeling of actual, other, generic: Most used Rich ERE trained classifiers
  - **BBN**: rules for actual vs. other
- Only Tinkerbell reports significant differences between languages
  - Used English system on machine translations of Spanish

Approaches to Hoppers Varied

… She will attend the conference. Next week's meeting …. →

Contact.Meet

** Participant, she=Marjorie Freedman, Other*

** Date, next week=W48-207, Other*

- Several relied on their event nugget co-reference
  - BUPT, CMU_CS (some runs)
- Tinkerbell trained classifiers to produce similarity scores of nuggets
- BBN used a sieve based approach

Evaluation Results

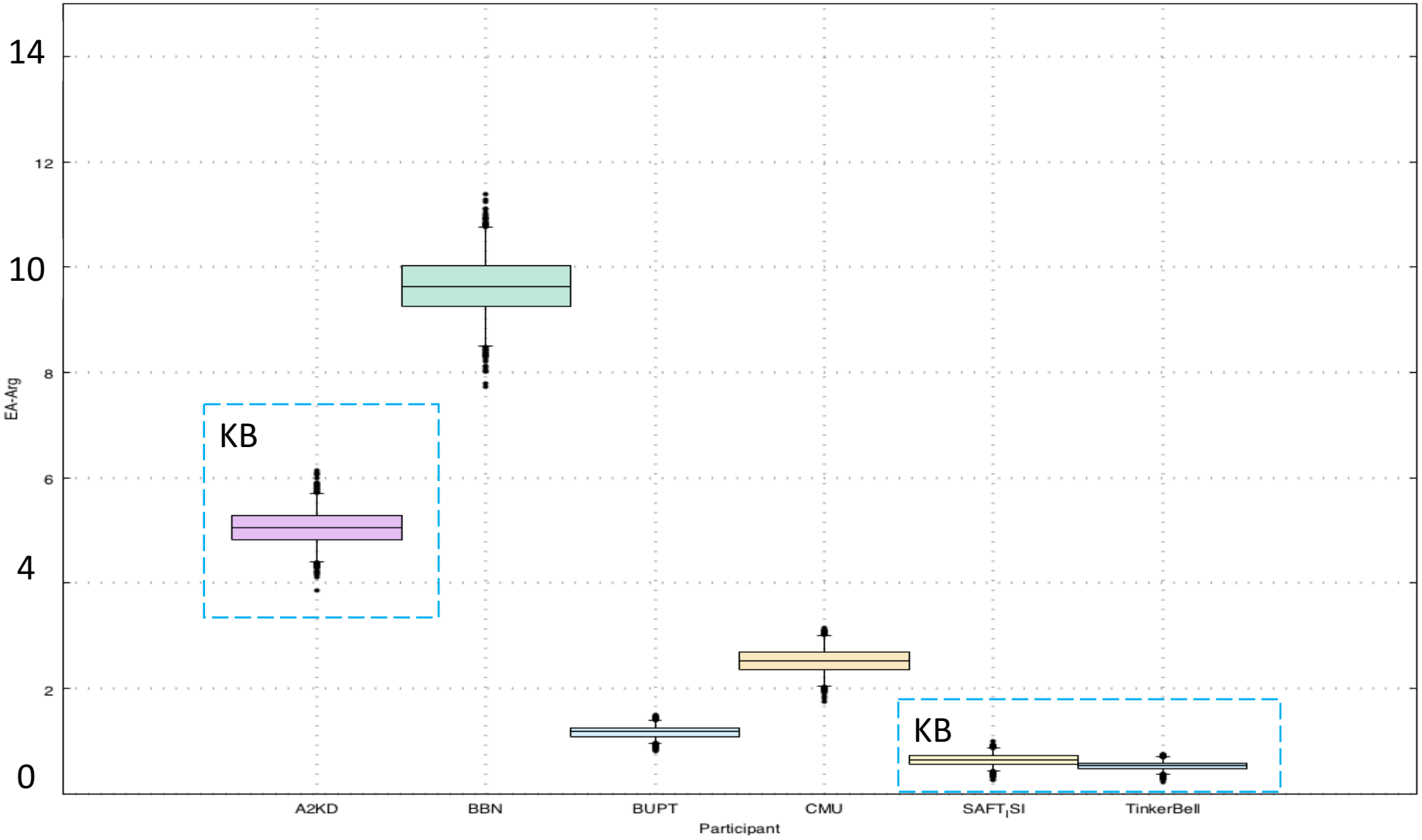
Argument Score

- Align *(EventSubtype, Role, Argument_Entity, Realis)* assertions with gold standard
  - Canonical Argument String serves as surrogate for Entity ID

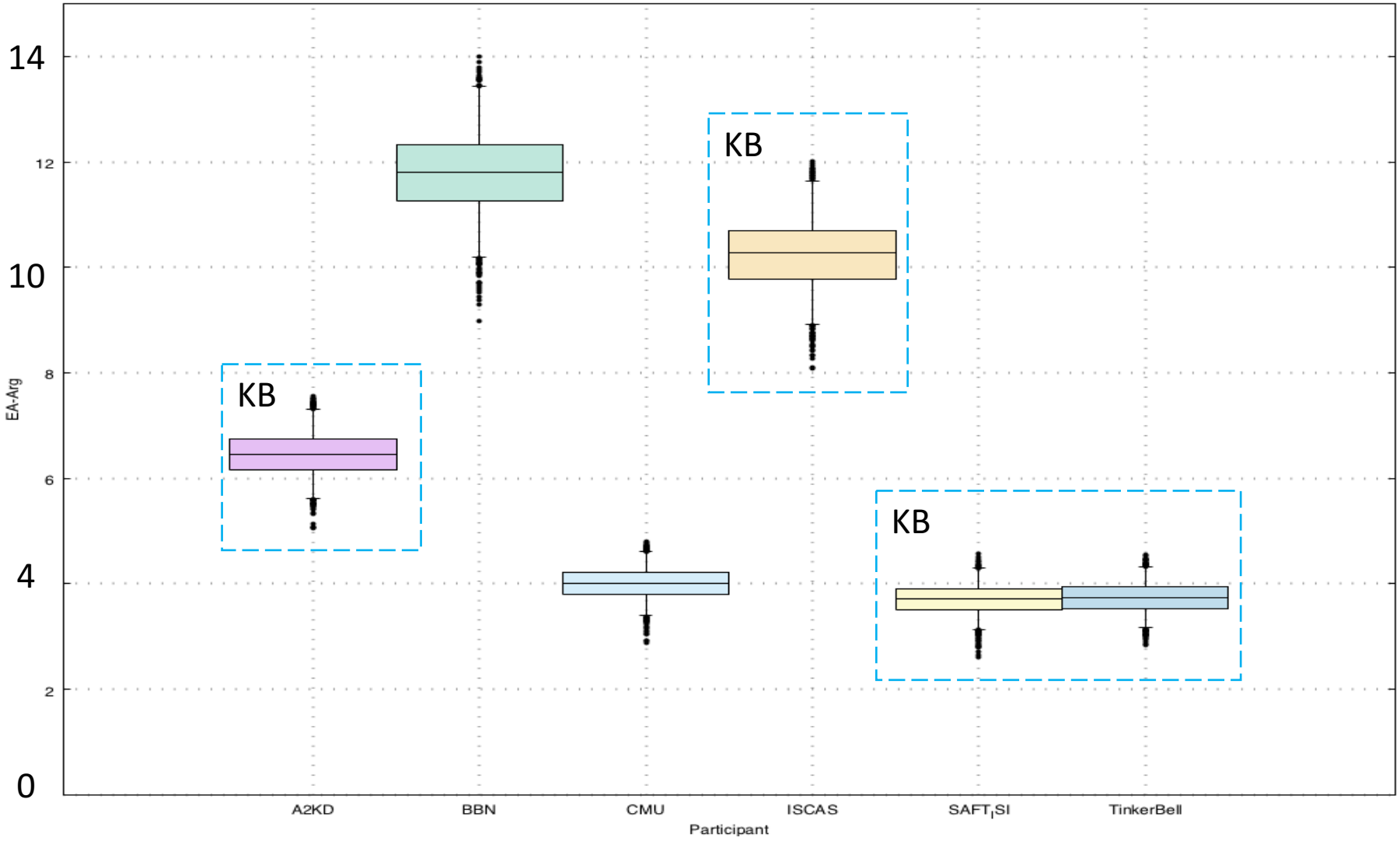
INJURE	VICTIM	At least six	Actual
INJURE	VICTIM	six people	Actual
INJURE	PLACE	Bahcelievler district	Actual
INJURE	PLACE	Istanbul	Actual
INJURE	DATE	Mon.(2006-02-13)	Actual
ATTACK	ATTACKER	TAK	Actual
ATTACK	TARGET	At least six	Actual
…	…	…	

- ArgScore: Error-based metric
  - Each document: $TP(d) - \beta FP(d)$
  - Over corpus: $\frac{1}{N}\sum_{d \in D}[max(0, arg(d))]$

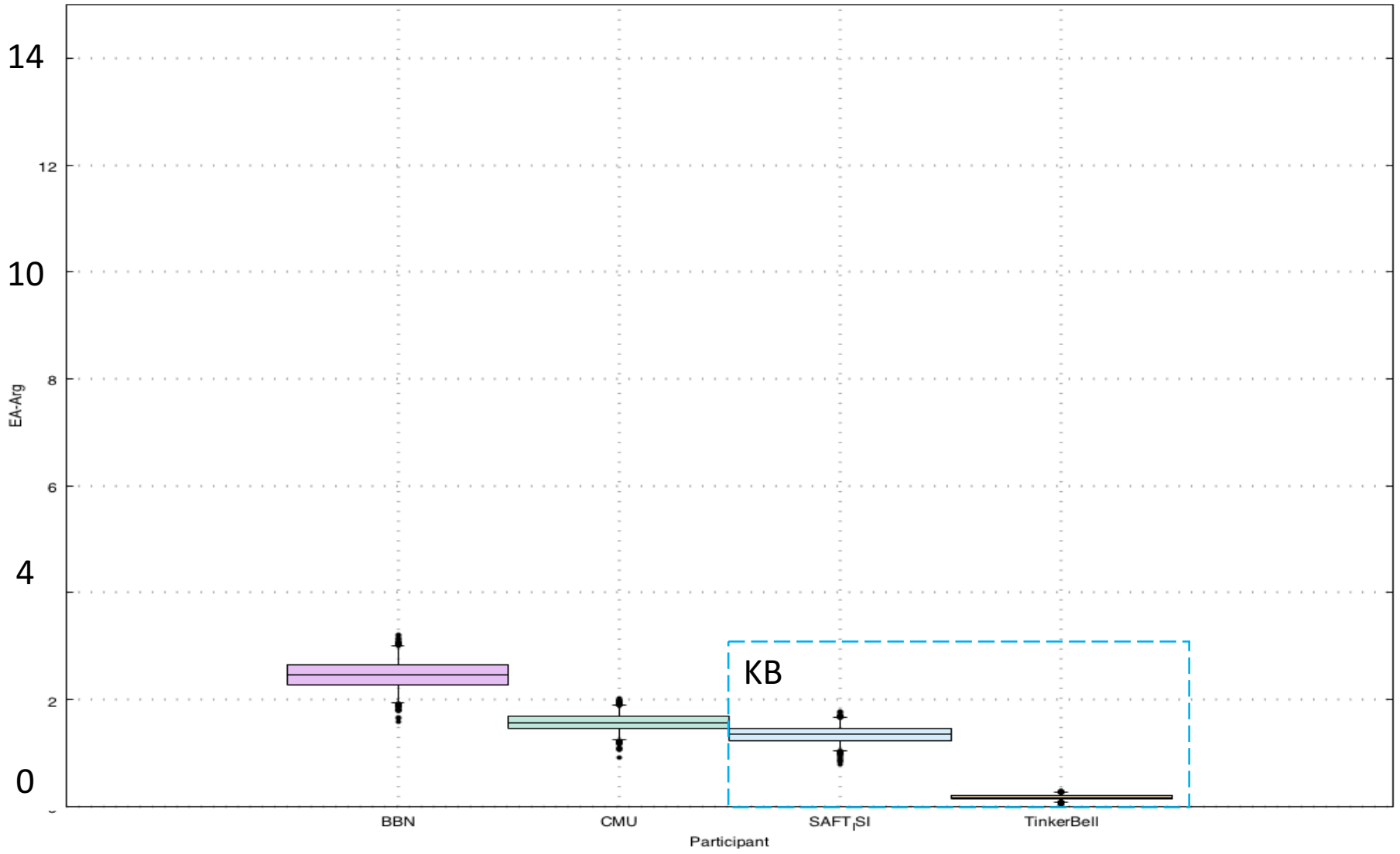
English Argument Scores



Chinese Argument Scores



Spanish Argument Scores



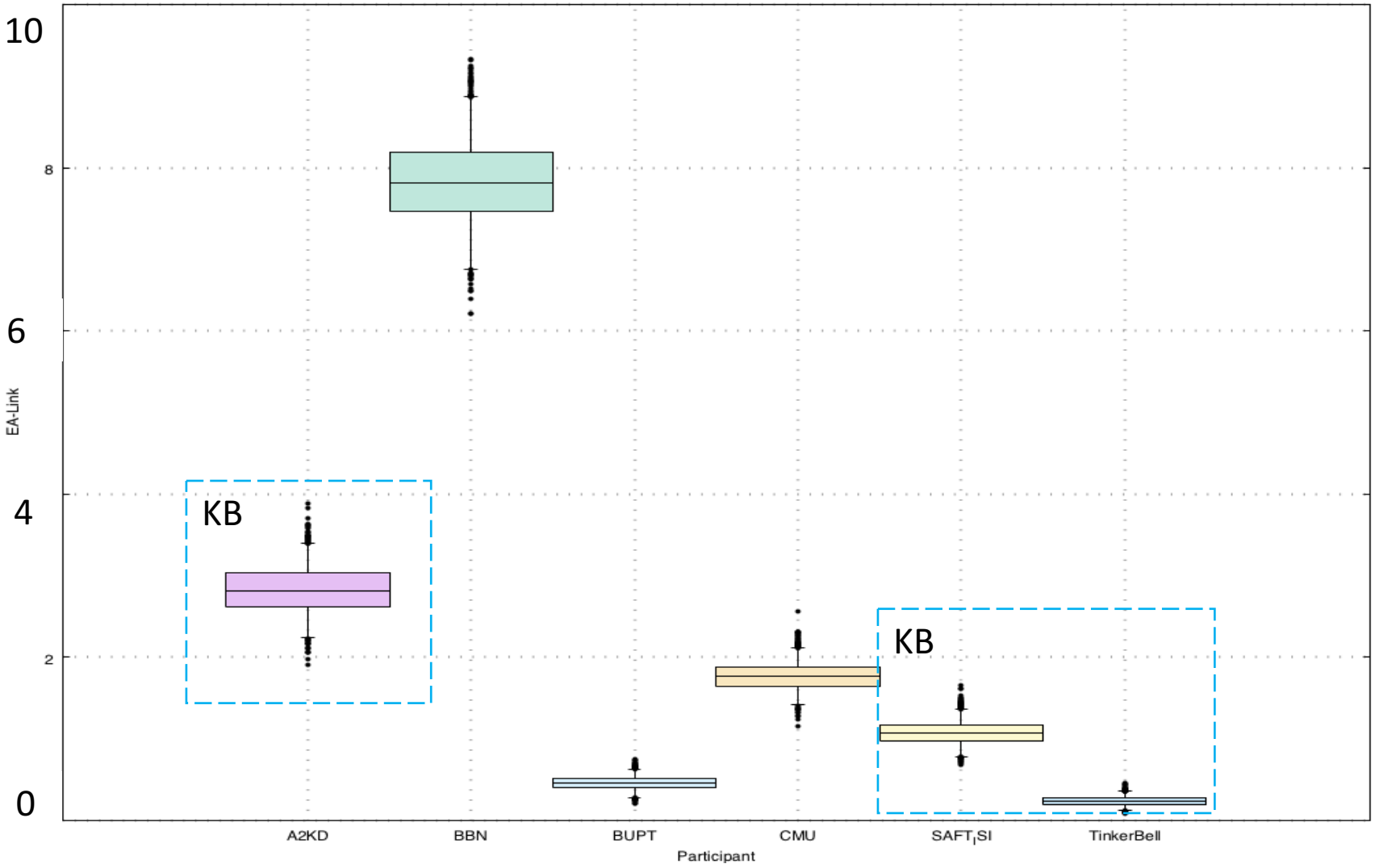
Linking (Hopper) Score

- Compare system hoppers with gold standard hoppers with B^3
  - Like argument score, measured at entity (and not mention) level
- Scoring of Hoppers
  - Ignores argument false positives
  - Limited by system recall

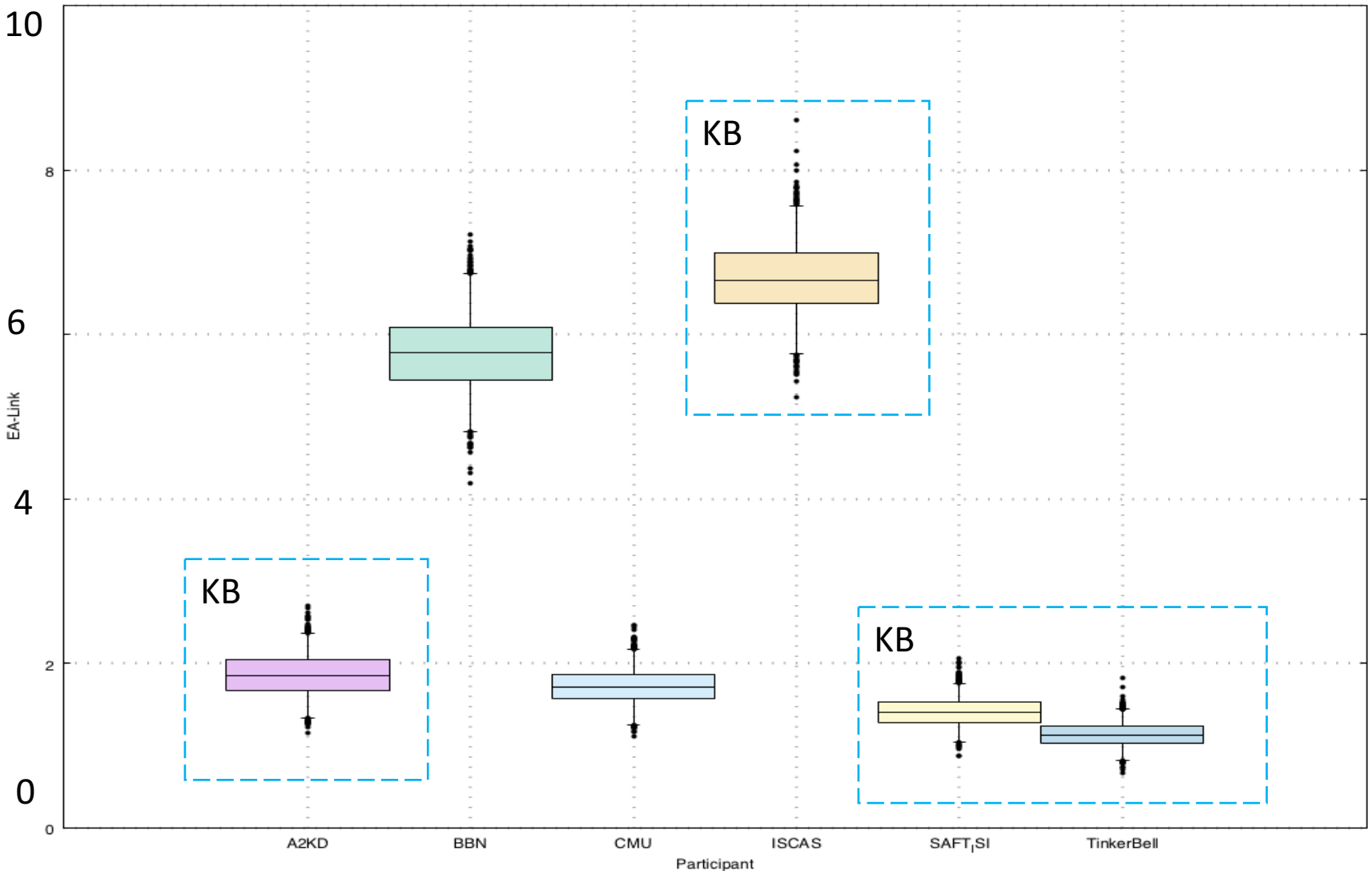
Event1	Role	Fillers
Life. Injure	Agent	TAK
	Victims	Six people 15 other people
	PLACE	the Bahcelievler district Istanbul An Istanbul supermarket
	DATE	Monday (2006-02-13)

Event2:	Role	Fillers
Conflict .Attack	ATTACKER	TAK
	TARGET	Six people 15 other people
	PLACE	the Bahcelievler district Istanbul An Istanbul supermarket
	DATE	Monday (2006-02-13)

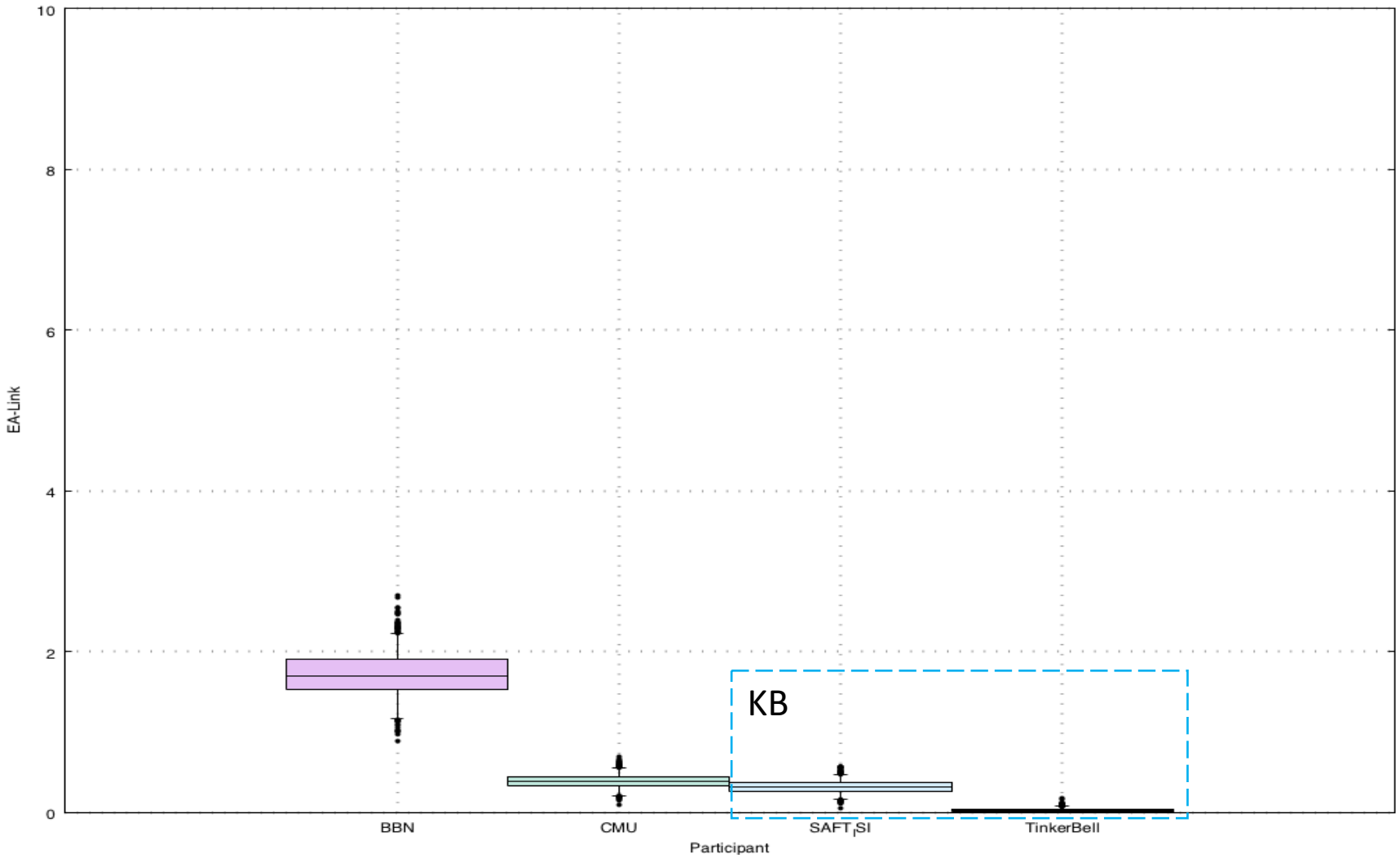
English Linking (Hopper) Scores



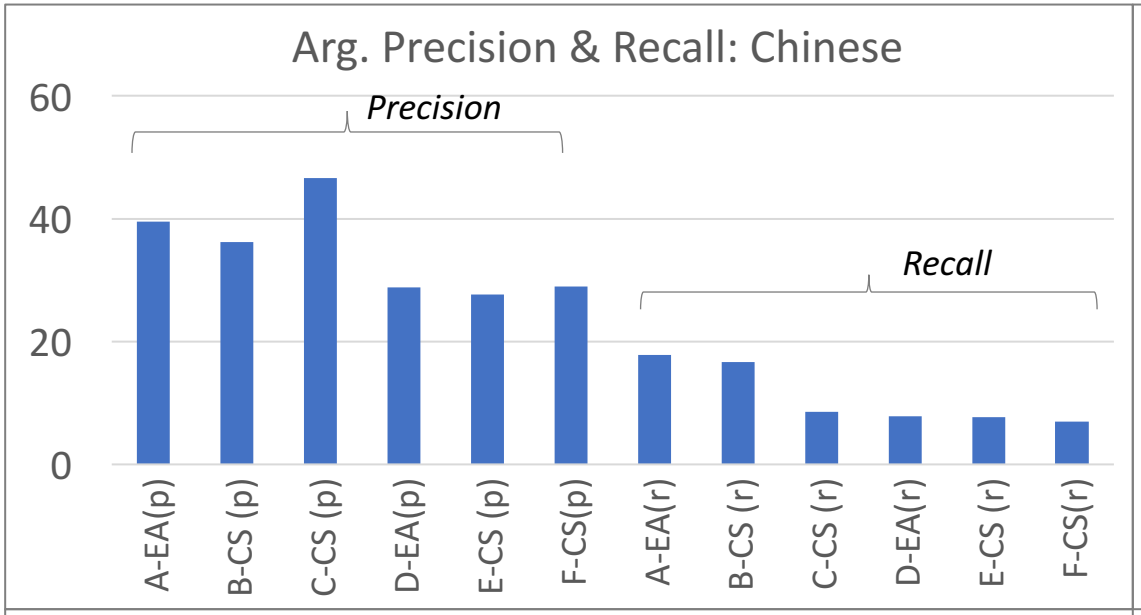
Chinese Linking (Hopper) Scoresß

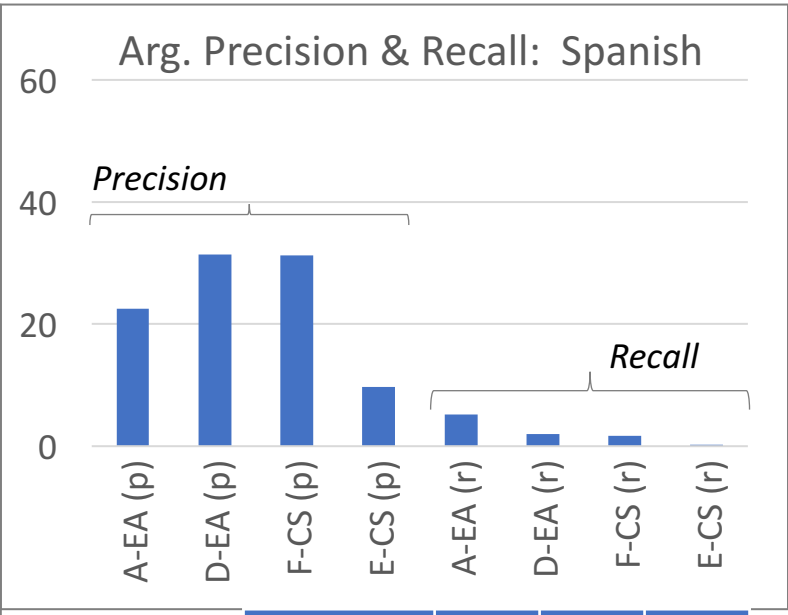


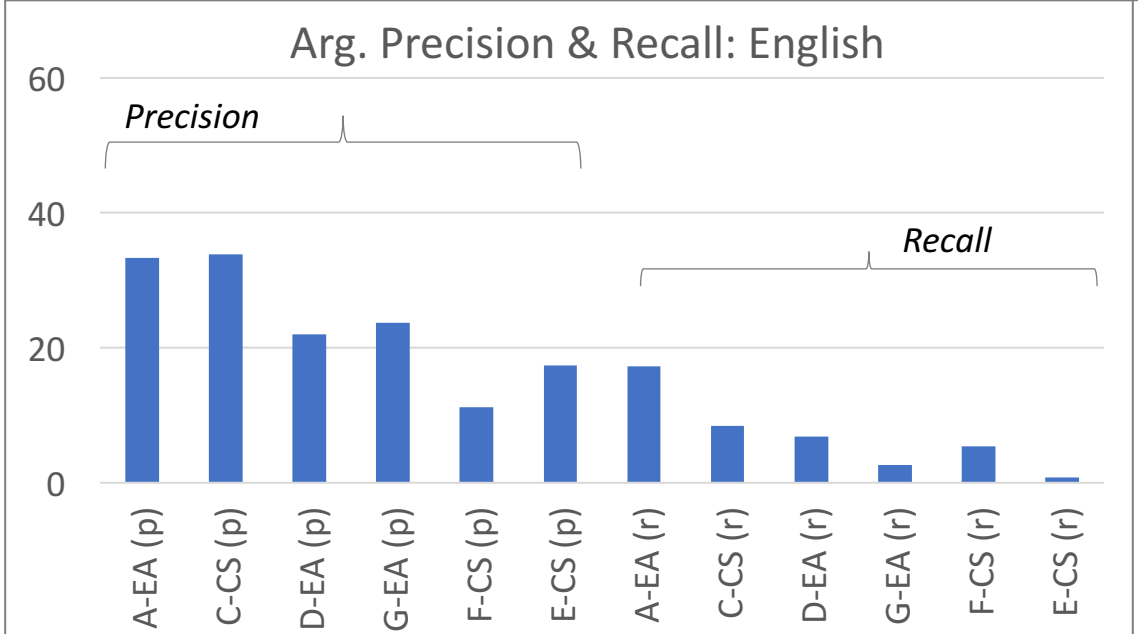
Spanish Linking (Hopper) Scores



Analysis of Argument Scores



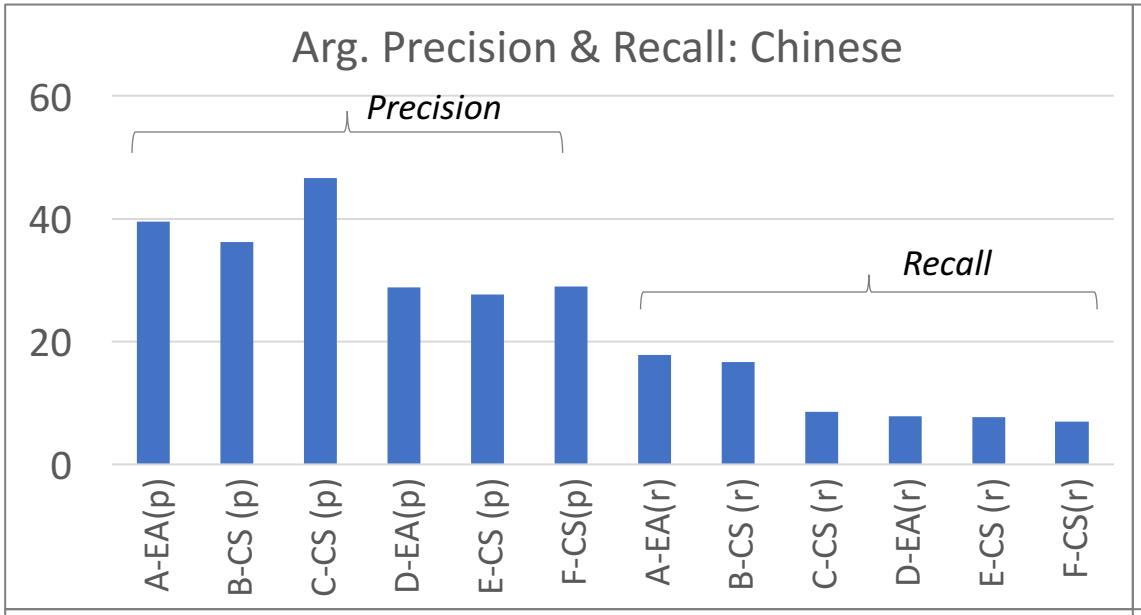


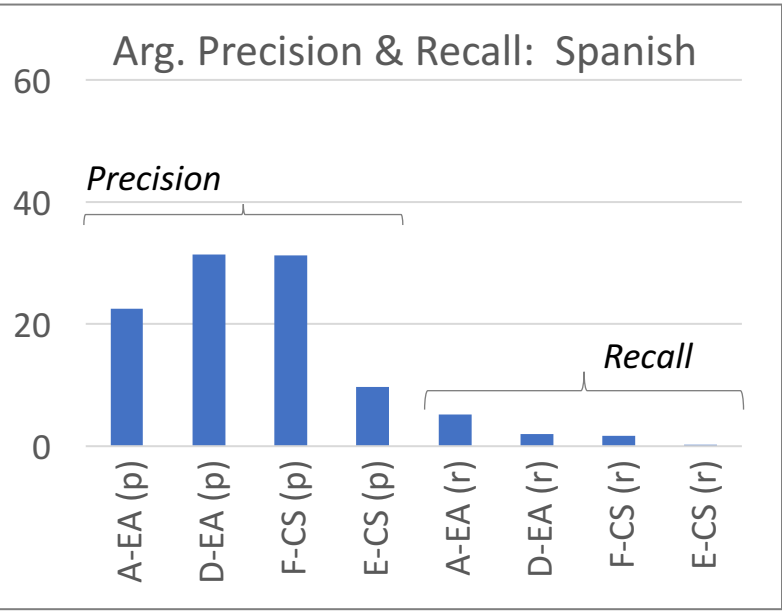


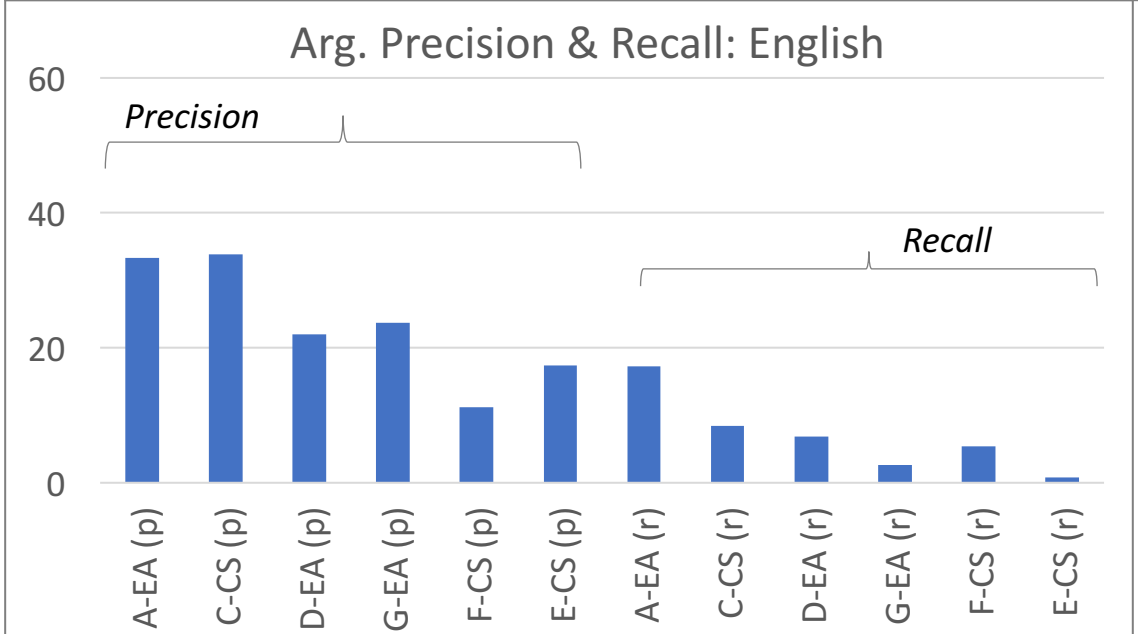
F1

	Ch	En	Sp
A-EA	***24***	***23***	***8***
B-CS	23	--	--
C-CS	14	13	--
D-EA	12	10	4
E-CS	12	2	0
F-CS	11	7	3
G-EA	--	5	--

Precision and Recall



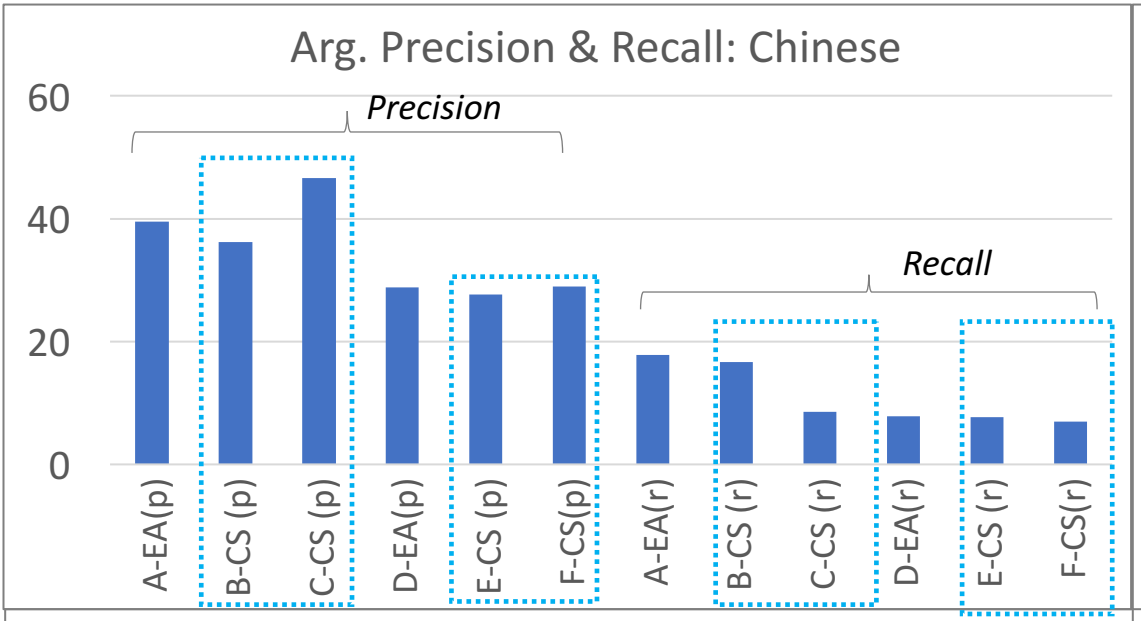


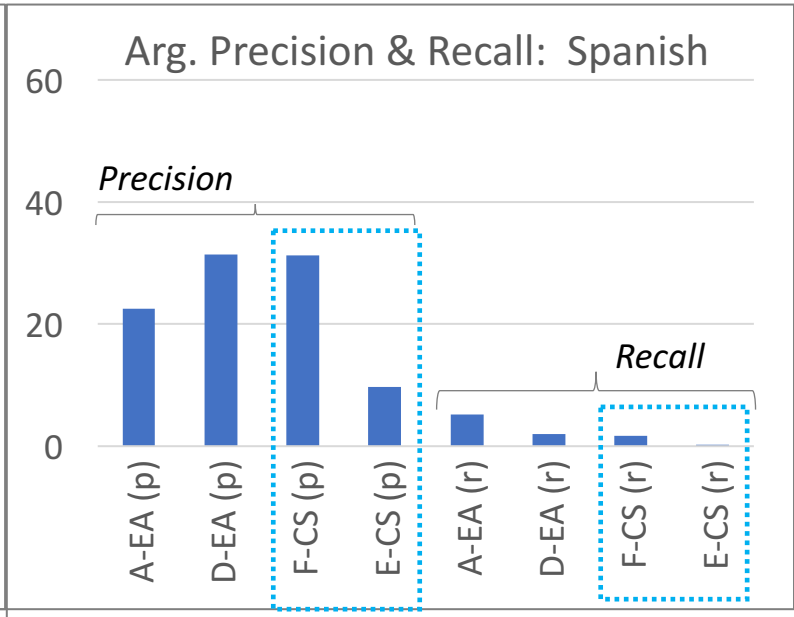


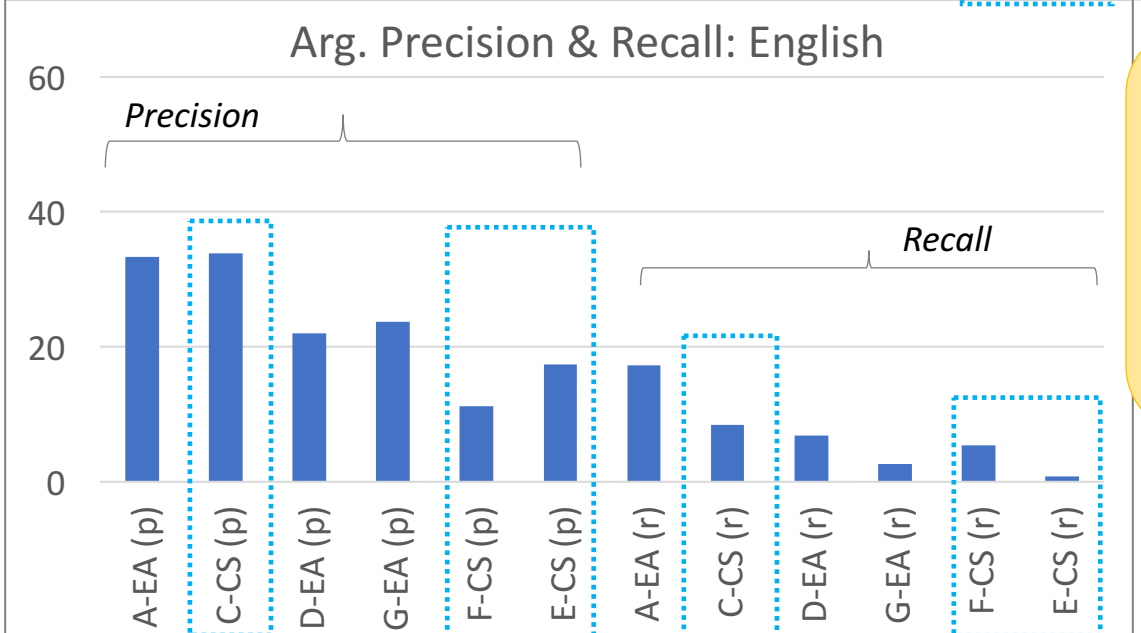
Recall lags precision

- For all languages
- For all systems

ColdStart++ vs. EAL Only



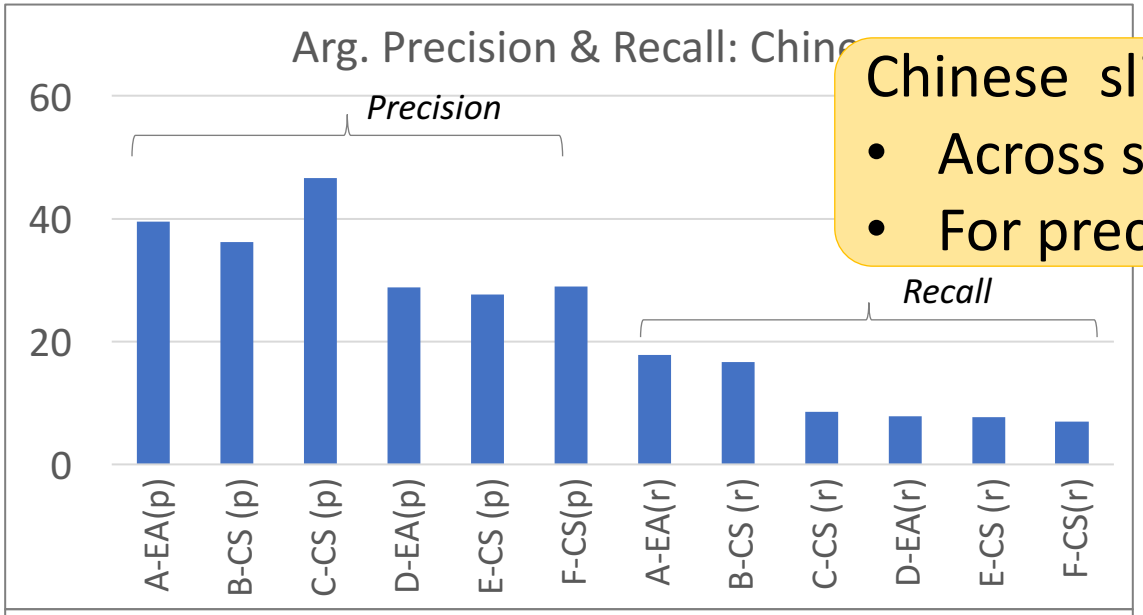


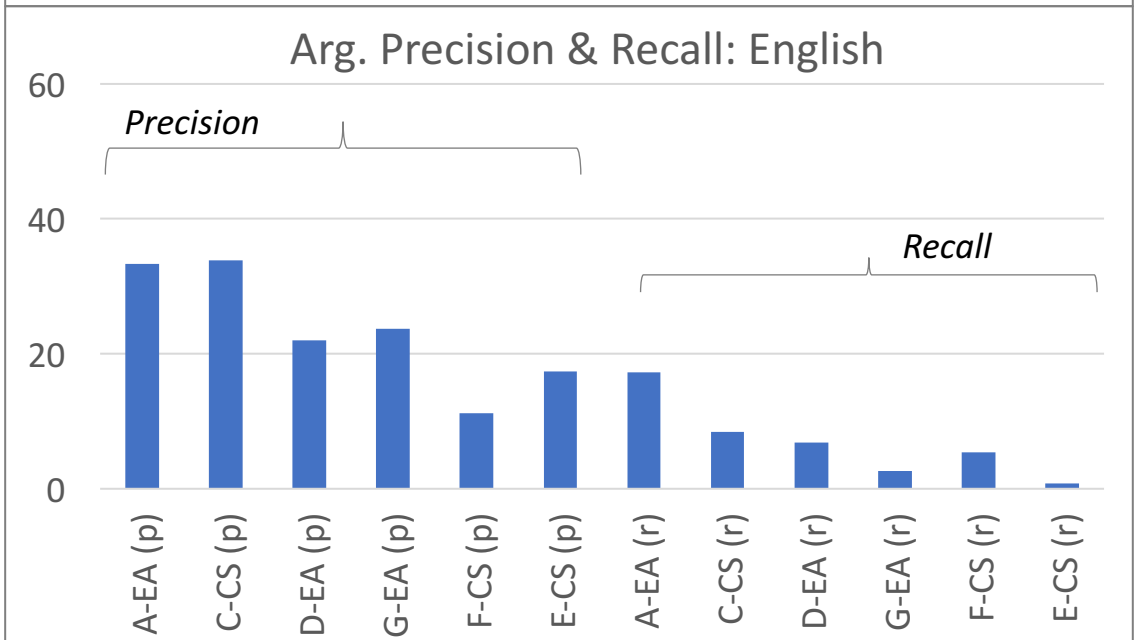


In general, EAL-only systems outperform ColdStart++

Why?
How can we better integrate the best EAL output into the KB?

Performance Across Languages (1)





Chinese slightly outperforms English

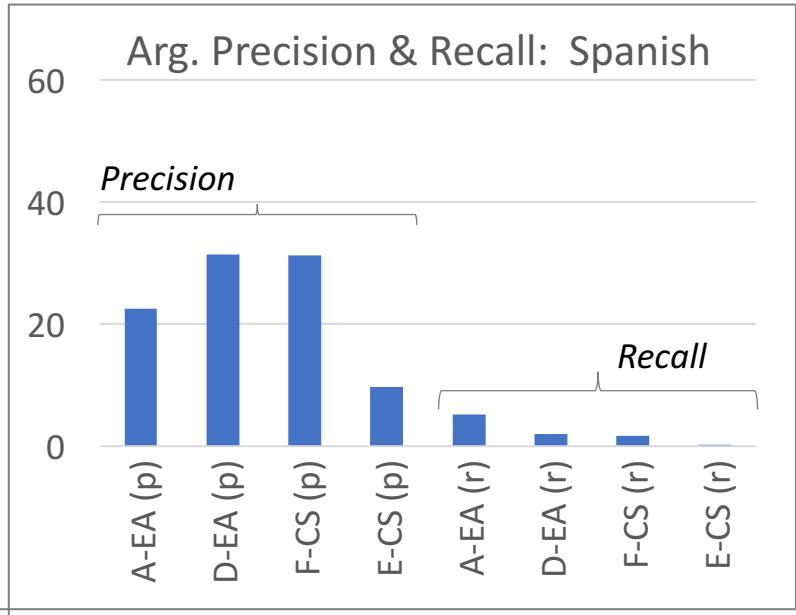
- Across systems
- For precision and recall

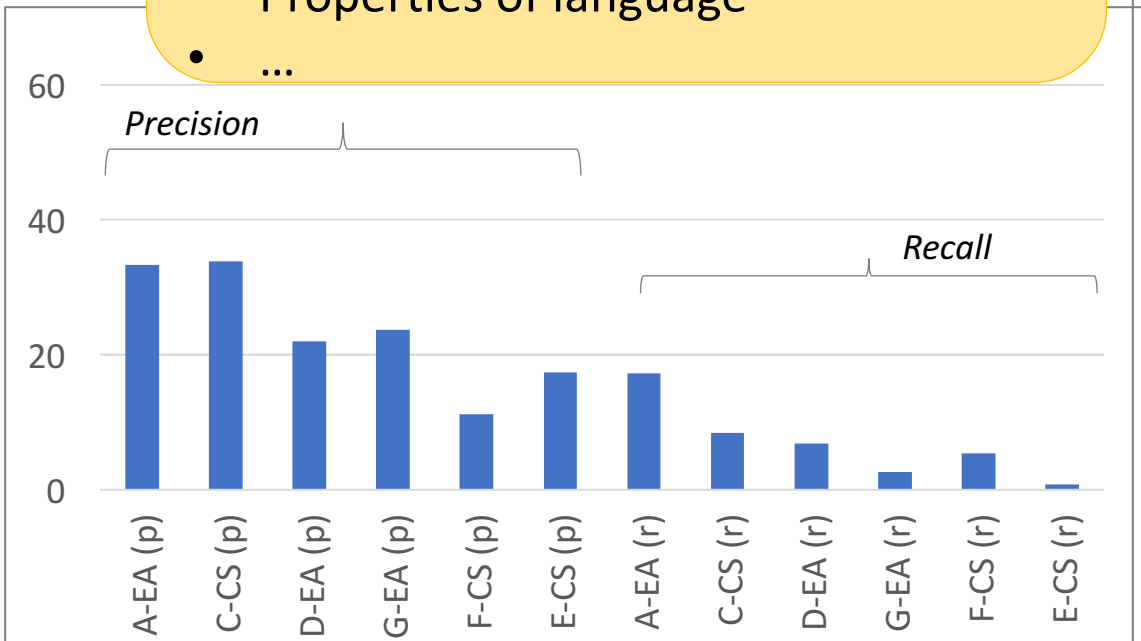
	Ch	En
A-EA	***24***	***23***
B-CS	23	--
C-CS	14	13
D-EA	12	10
E-CS	12	2
F-CS	11	7
G-EA	--	5

Performance Across Languages (2)

Spanish performance lags English
- Across systems
- Especially for recall

Why?
- Less training data
- Less accurate linguistic processing (parsing, coreference, etc.)
- Characteristic of test set
- Properties of language
- …





	En	Sp
A-EA	***23***	***8***
C-CS	13	--
D-EA	10	4
E-CS	2	0
F-CS	7	3
G-EA	5	--

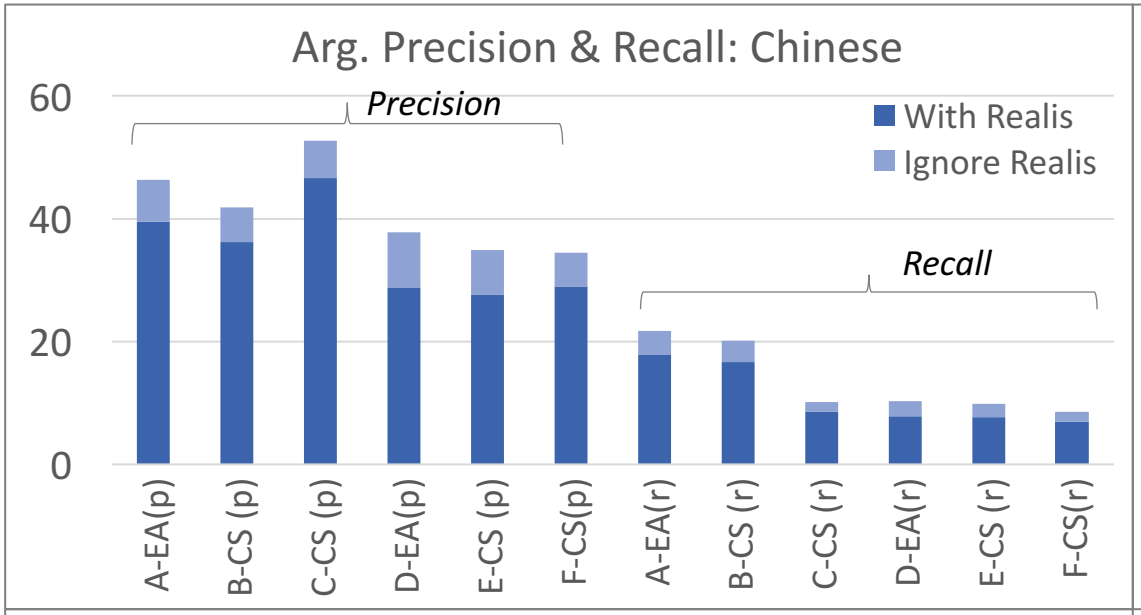
Performance Across Languages (3)

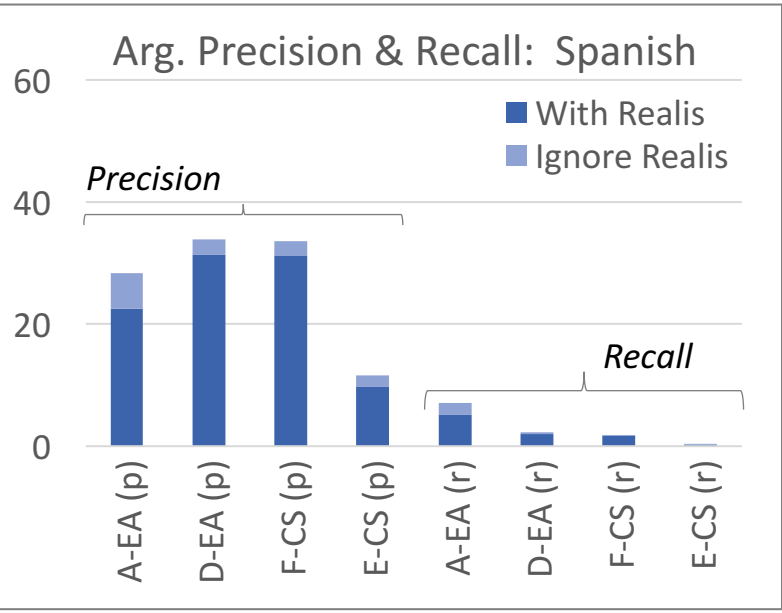
Argument F1

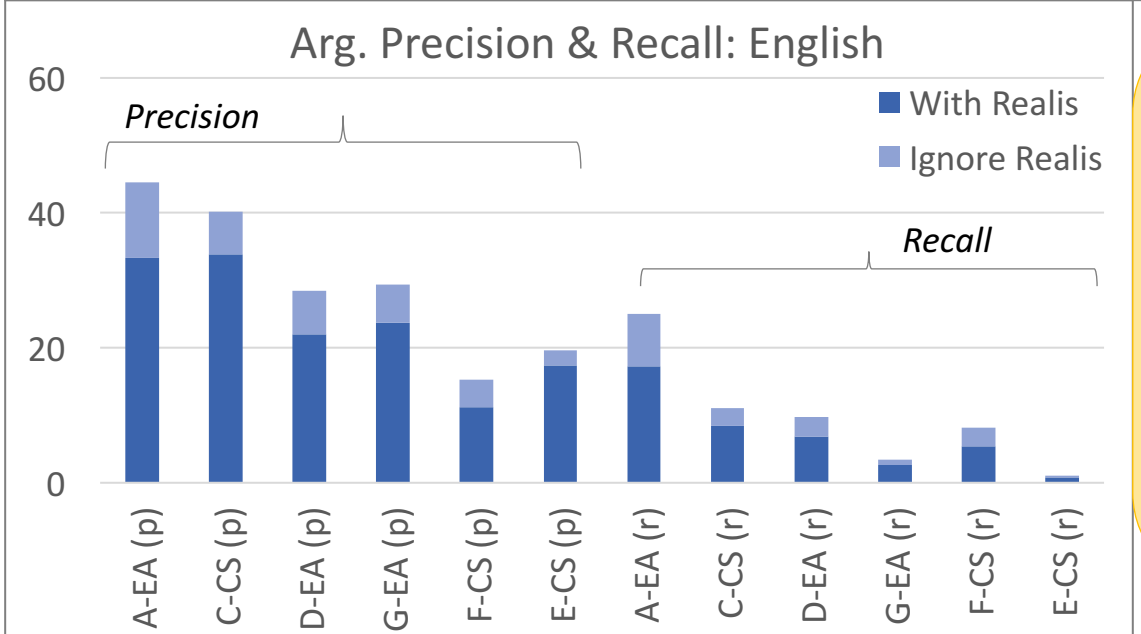
	Ch	En	Sp
A-EA	***24***	***23***	***8***
B-CS	23	--	--
C-CS	14	13	--
D-EA	12	10	4
E-CS	12	2	0
F-CS	11	7	3
G-EA	--	5	--

- System rank is relatively constant across languages
- At current performance levels, techniques transfer relatively well between languages
- But, current performance levels are low in absolute terms

Actual vs. Other vs. Generic







Ignoring realis distinction *(actual, generic, other)*

- Improves precision & recall
- Improves performance in all languages
- But, absolute performance remains low (i.e. F1: ~30 for top performing EN & CH)

What's Next?

- 2018 is TBD
- 2014-2017 EAL tasks have resulted in
  - More training data (RichERE)
  - A scoring package that measure event argument performance at the level of a KB assertion
    - https://github.com/isi-nlp/tac-kbp-eal
  - Two shared tests sets
- What would help improve system performance?
- Are people interested in this task outside of TAC
  - Would it help to share 2016 and 2017 system output for future comparison?
    - Hosted with scorer?