Overview of the KBP 2015 Slot Filler Validation Track

Hoa Trang Dang
National Institute of Standards and Technology

Slot Filler Validation (SFV)

- Track Goals
  - Allow teams without a full slot-filling system to participate in KBP, focus on SF answer validation rather than IR, IE, EDL, etc.
  - Evaluate the contribution of RTE systems on KBP slot-filling
  - Allow teams to experiment with system voting and ensembling
- Piggy back off of resources developed for and by KBP [Cold Start] Slot Filling
- Task and evaluation metrics depend on use case and availability of additional information about candidate fillers
  - RTE: correctness of candidate slot filler is judged in isolation – no knowledge of who proposed the candidate slot filler. Generally requires going back to the source documents
  - SFV: candidate slot fillers grouped according to which system propose the slot filler – leverage wisdom of the crowd

SFV 2015

- SFV input:
  - All KBP 2015 CS Slot Filling input (slot definitions, CSSF queries, source documents)
  - Anonymized individual CS KB/SF runs
    - SFV2015_KB_12_5
    - SFV2015_KB_2_1
    - SFV2015_SF_2_1
  - System profile for each CS run (“are the confidence values meaningful?”)
  - Preliminary assessment of ~10% of CSSF queries (164 / 1983)
  - Mapping to real team names *(extra)*
    - SFV2015_KB_12 = “BBN”
    - SFV2015_KB_2 = “Stanford KB”
    - SFV2015_SF_2 = “Stanford SF”
- SFV output: Binary classification of each candidate slot filler in each CS run (-1/+1 : Exclude/Include slot filler)

Task 1: SFV Filtering Task

- Apply SFV filter to set of original CS runs to produce a filtered version of each original CS run.

- *Can only improve Precision, not Recall, of individual CS runs*

- Score each original and filtered CS run with Cold Start scorer, and report change in F1

- Final SFV Filtering score = mean change in F1, over all CS runs
  - How much can you improve an individual CS run, on average?

Task 2: SFV Ensemble Task

- Apply SFV filter to set of original CS runs to produce a single ensemble CS run

- *Possible to improve both Precision and Recall over original CS runs*

- Score ensemble CS run with Cold Start scorer

- Final SFV Ensemble score = F1 of the ensemble run

Applying Cold Start scorer in SFV

- CS scorer penalizes a CS run for returning multiple slot fillers that are duplicates (refer to the same entity, concept, etc.).
  - SFV must optimally remove duplicate “Correct” candidate slot fillers within a CS run and (for ensemble) across the set of CS runs.
- Identifying that different Cold Start entry points are for the same entity is currently outside the scope of SFV
- SFV evaluation focuses on *micro-average* Cold Start scores -- each correct slot filling answer (equivalence class) is weighted evenly.

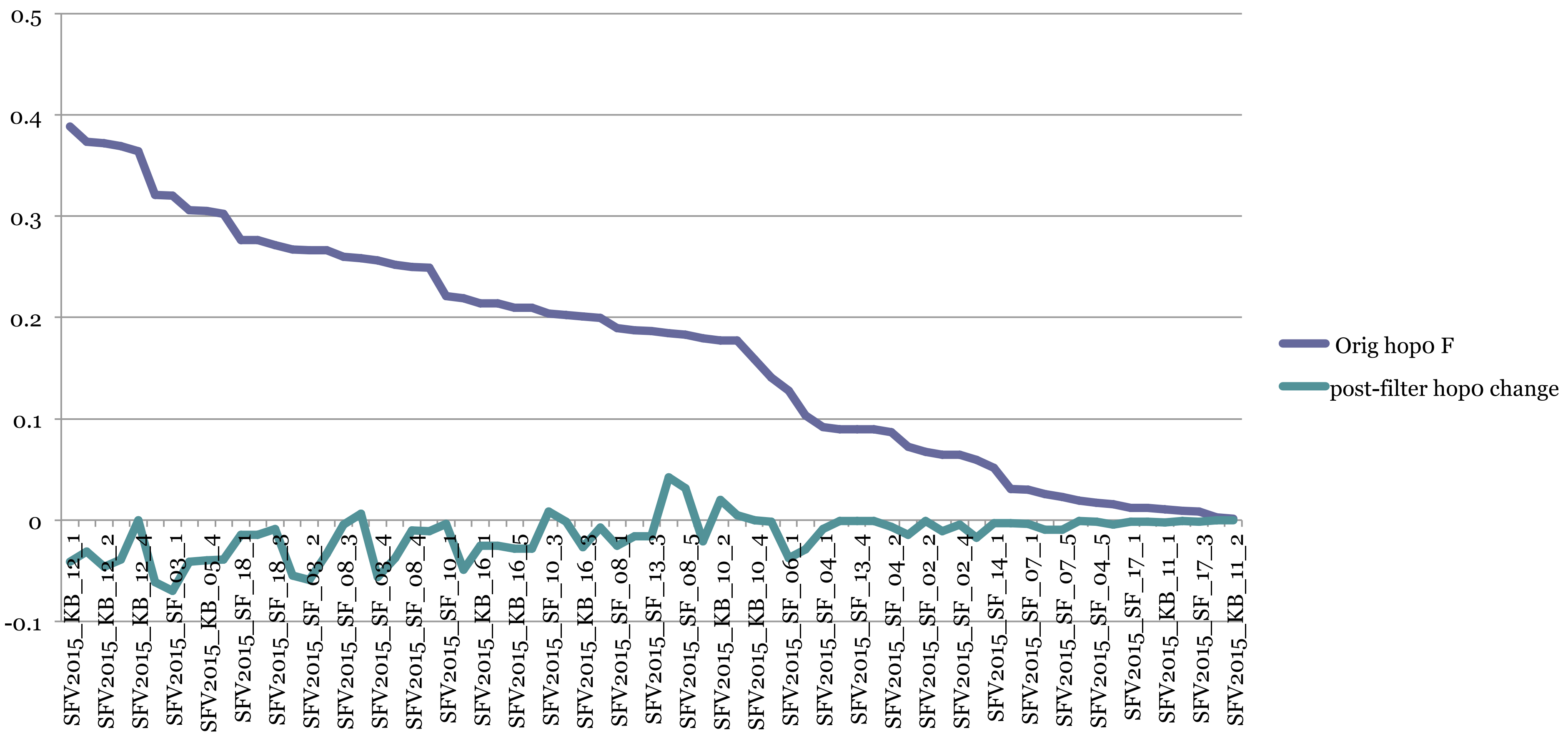
- Score only on the 90% of CSSF queries that did *not* have preliminary assessments released as part of the SFV input

SFV 2015 Participants

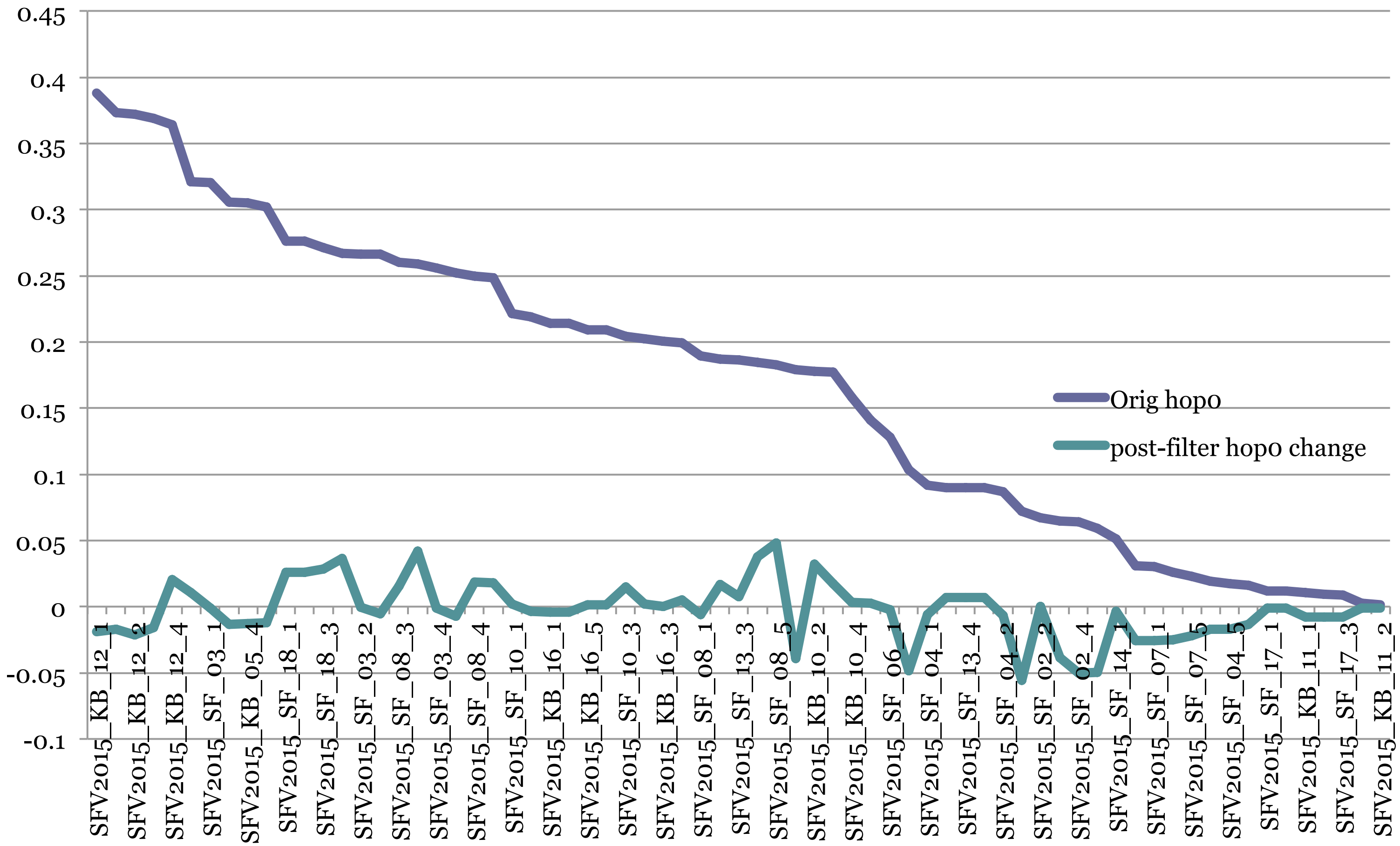
Team	Organization	Confidence	Assessment
* gator_dsr	University of Florida	Yes	Yes
jhuapl	Johns Hopkins University Applied Physics Laboratory	Yes	Yes
RPI_BLENDER	Rensselaer Polytechnic Institute	No	Yes
UI_CCG	University of Illinois Urbana Champaign	No	Yes
* UTAustin	University of Texas at Austin	Yes	Yes

* SFV team was provided with real identity of Cold Start teams (build on UTAustin work on supervised ensembling)

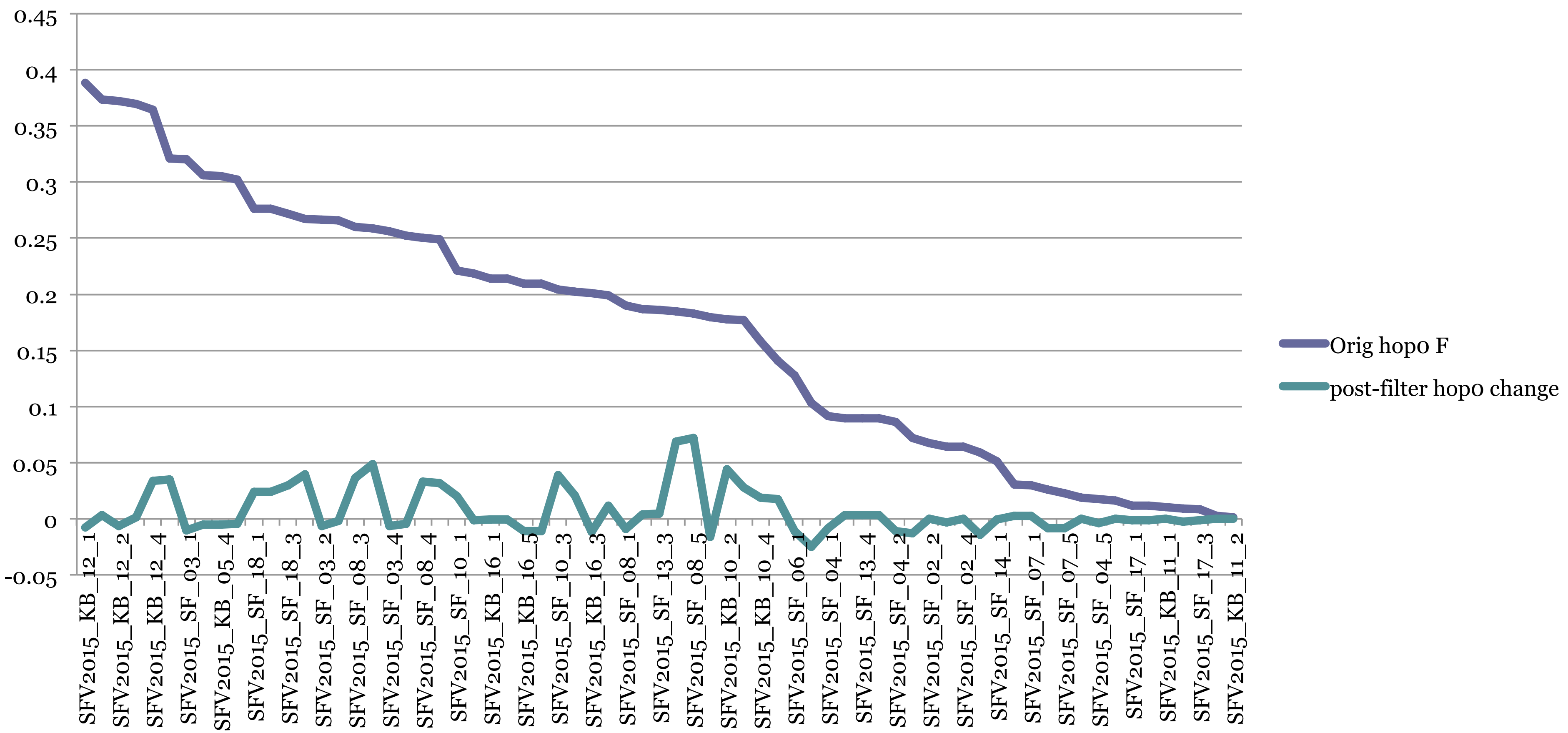
jhuapl1 filter (cssf micro-average)



RPI_BLENDER1 filter (cssf micro-average)



gator_dsr3 filter (cssf micro-average)



Top 20 CSSF runs (cssf micro-average)

SFV run	CSSF run	Hop0 F1
gator_dsr2	ensemble	0.45
gator_dsr3	ensemble	0.44
gator_dsr1	ensemble	0.44
gator_dsr3	SFV2015_KB_12_4.filtered	0.4
gator_dsr2	SFV2015_KB_12_4.filtered	0.4
UI_CCG1	SFV2015_KB_12_1.filtered	0.39
--	SFV2015_KB_12_1	0.39
RPI_BLENDER2	SFV2015_KB_12_4.filtered	0.38
RPI_BLENDER1	SFV2015_KB_12_4.filtered	0.38
gator_dsr3	SFV2015_KB_12_1.filtered	0.38
gator_dsr2	SFV2015_KB_12_1.filtered	0.38
gator_dsr3	SFV2015_KB_12_3.filtered	0.38
gator_dsr2	SFV2015_KB_12_3.filtered	0.38
UI_CCG1	SFV2015_KB_12_3.filtered	0.37
--	SFV2015_KB_12_3	0.37
UI_CCG1	SFV2015_KB_12_2.filtered	0.37
--	SFV2015_KB_12_2	0.37
gator_dsr3	SFV2015_KB_12_5.filtered	0.37
gator_dsr2	SFV2015_KB_12_5.filtered	0.37
UI_CCG1	SFV2015_KB_12_5.filtered	0.37

Conclusion

- SFV is able to improve on state-of-the art Cold Start 2015 KB/SF systems
- Difficult to optimize SFV filter to help all/most Cold Start runs
- “partial preliminary assessments” provide only weak indication of performance of each Cold Start run.
- Real Cold Start team IDs help significantly – leverage past results for teams that participated in past SF tracks
- *Should we always provide real CS team IDs in future?*