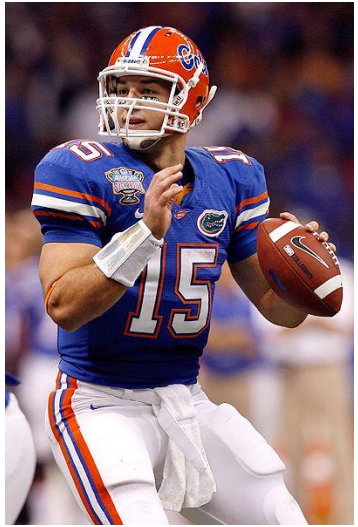
University of Florida DSR Lab System for KBP Slot Filler Validation 2015

Miguel Rodriguez, Sean Goldberg, Daisy Wang

DSR @ UF
Data Science Research

Slot Filler Validation



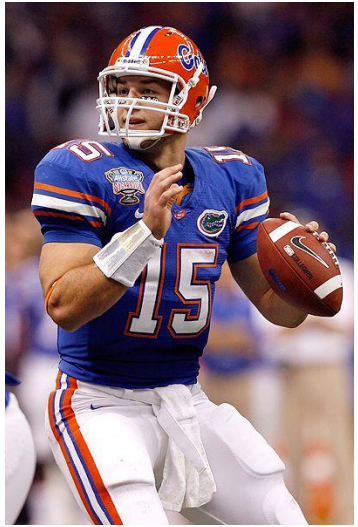
Tim Tebow

gpe:schools_atended

Bristol Central High School
New England Patriots
University of Florida
University of Connecticut
ABC News

DSR @ UF
Data Science Research

Slot Filler Validation



Tim Tebow

gpe:schools_atended

	Truth
Bristol Central High School	T
New England Patriots	F
University of Florida	T
University of Connecticut	F
ABC News	F

DSR @ UF
Data Science Research

Slot Filler Validation

org:subsidiaries

	Truth
Survey Research Center	T
Florida Museum of Natural History	T
Smithsonian Tropical Research Institute	F

DSR @ UF
Data Science Research

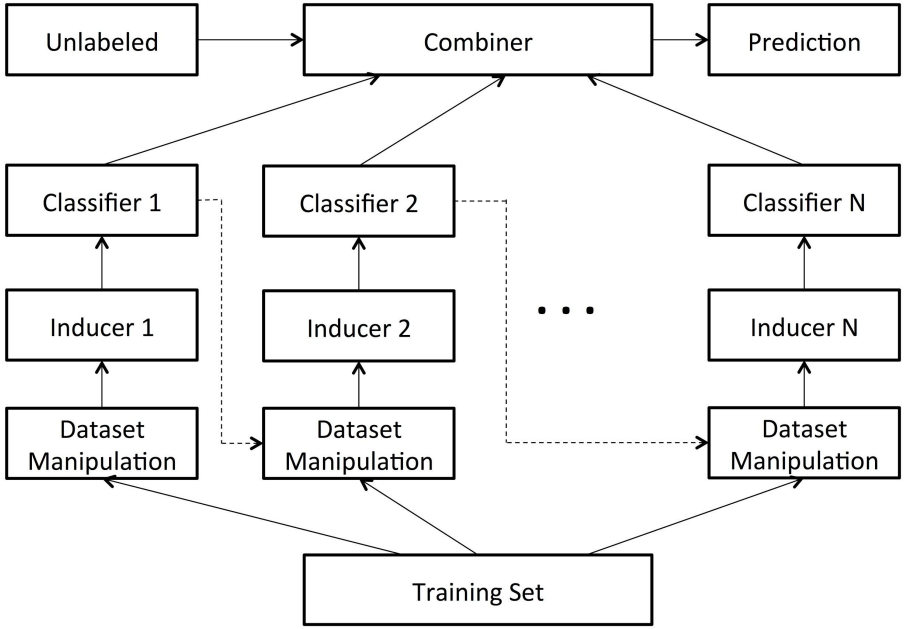
Slot Filler Validation - Classification

- Slot Filler Validation is a binary classification task
  - Given a set of queries consisting of tuples of the form <entity, slot>
  - And a set of Slot Fillers for each query
  - Determine if a slot filler is True or False

Slot Filler Validation - Classification

- Slot Filler Validation is a binary classification task
  - Given a set of queries consisting of tuples of the form <entity, slot>
  - And a set of Slot Fillers for each query
  - Determine if a slot filler is True or False
- A CSSF output is the output of such classifier
  - Ideal for ensemble classification
  - Aggregate the output of multiple classifiers
  - Outperform the original ones

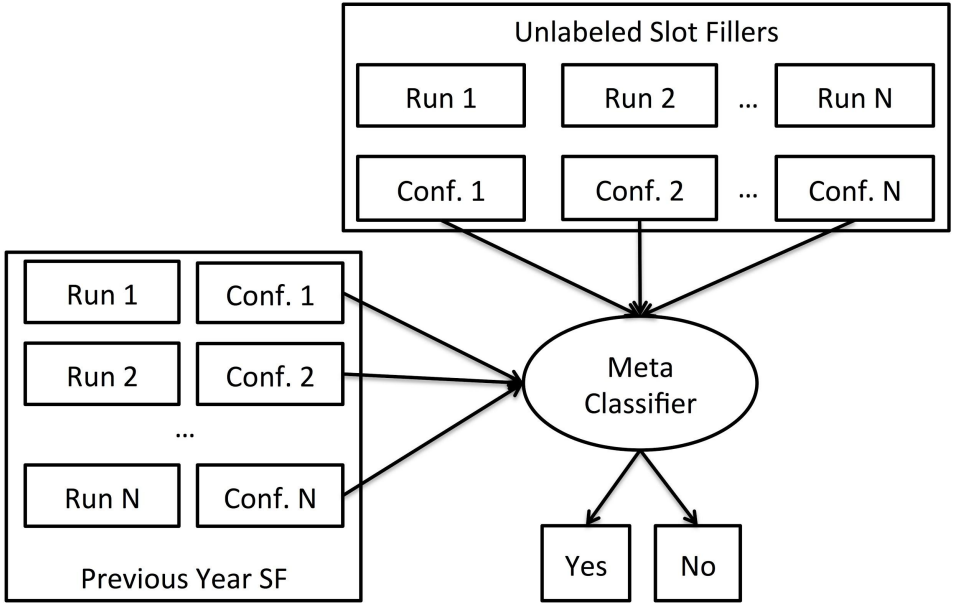
Ensemble Classification



- Ensemble methods have two main parts
  - **Inducer**: Selects the training data for each individual classifier
  - **Combiner**: takes the output of each classifier and combine them to formulate a final prediction

DSR @ UF
Data Science Research

Stacked Ensemble



Meta-level classifier that takes the output of other models as input and estimate their weights

Vidhoon Viswanathan, Nazneen Fatema Rajani, and Yinon Bentor Raymond J Mooney. 2015. Stacked ensembles of information extractors for knowledge-base population. In Proceedings of the 53rd annual meeting on association for computational linguistics. Association for Computational Linguistics

DSR @ UF
Data Science Research

Stacked Ensemble

- Requires labeled data
  - Available from 2013 and 2014 SF and SFV
- Training Strategy
  - Learn from previous year performance
  - 2013-2014: 7 teams
  - 2014: 12 teams

Stacked Ensemble

- Requires labeled data
  - Available from 2013 and 2014 SF and SFV
- Training Strategy
  - Learn from previous year performance
  - 2013-2014: 7 teams
  - 2014: 12 teams
- All runs that can not be fit into the classifier are discarded!
  - Leave out extra evidence
  - ... From potentially well ranked systems

DSR @ UF
Data Science Research

Stacked Ensemble - not enough!

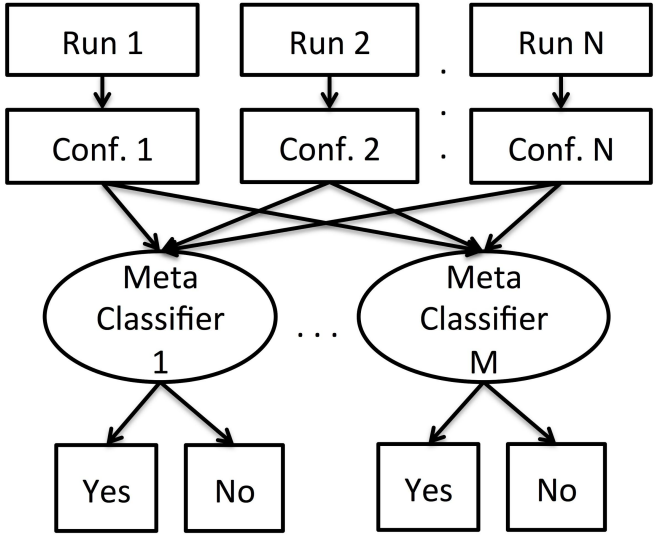
Rank	TEAM ID	0-HOP F1	1-HOP F1	ALL F1
9	SFV2015_SF_03_1	0.3457	0.1154	0.2718
14	SFV2015_KB_16_2	0.2633	0.1655	0.2247
16	SFV2015_SF_18_1	0.292	0.0972	0.2245
24	SFV2015_SF_08_4	0.2669	0.0976	0.2102
31	SFV2015_SF_02_1	0.1883	0.1299	0.1649
34	SFV2015_SF_06_1	0.2351	0	0.1595

Rank	TEAM ID	0-HOP F1	1-HOP F1	ALL F1
39	SFV2015_KB_10_1	0.1834	0.0952	0.1474
45	SFV2015_KB_09_1	0.0965	0.0791	0.0899
47	SFV2015_SF_13_2	0.1225	0	0.0892
56	SFV2015_SF_07_1	0.0512	0	0.0353
63	SFV2015_KB_11_1	0.019	0	0.0121
64	SFV2015_SF_17_1	0.019	0	0.0121

F1 score ranking of 2014-2015 teams.

DSR @ UF
Data Science Research

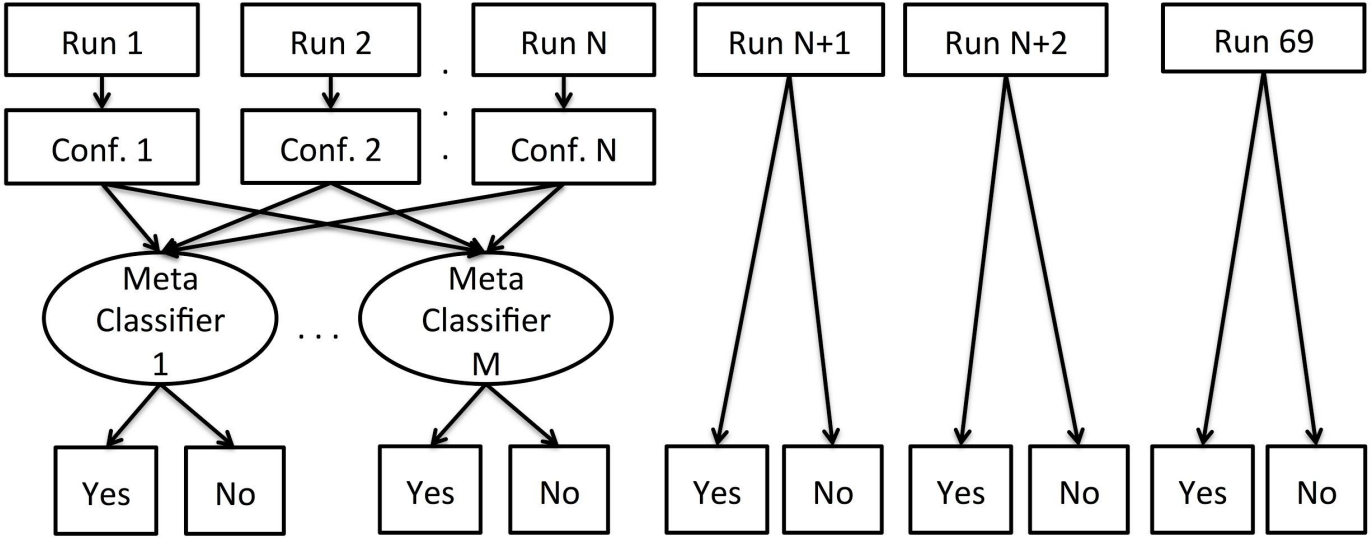
Consensus Maximization Fusion



Augment stacked ensemble model by adding more meta-classifiers

DSR @ UF
Data Science Research

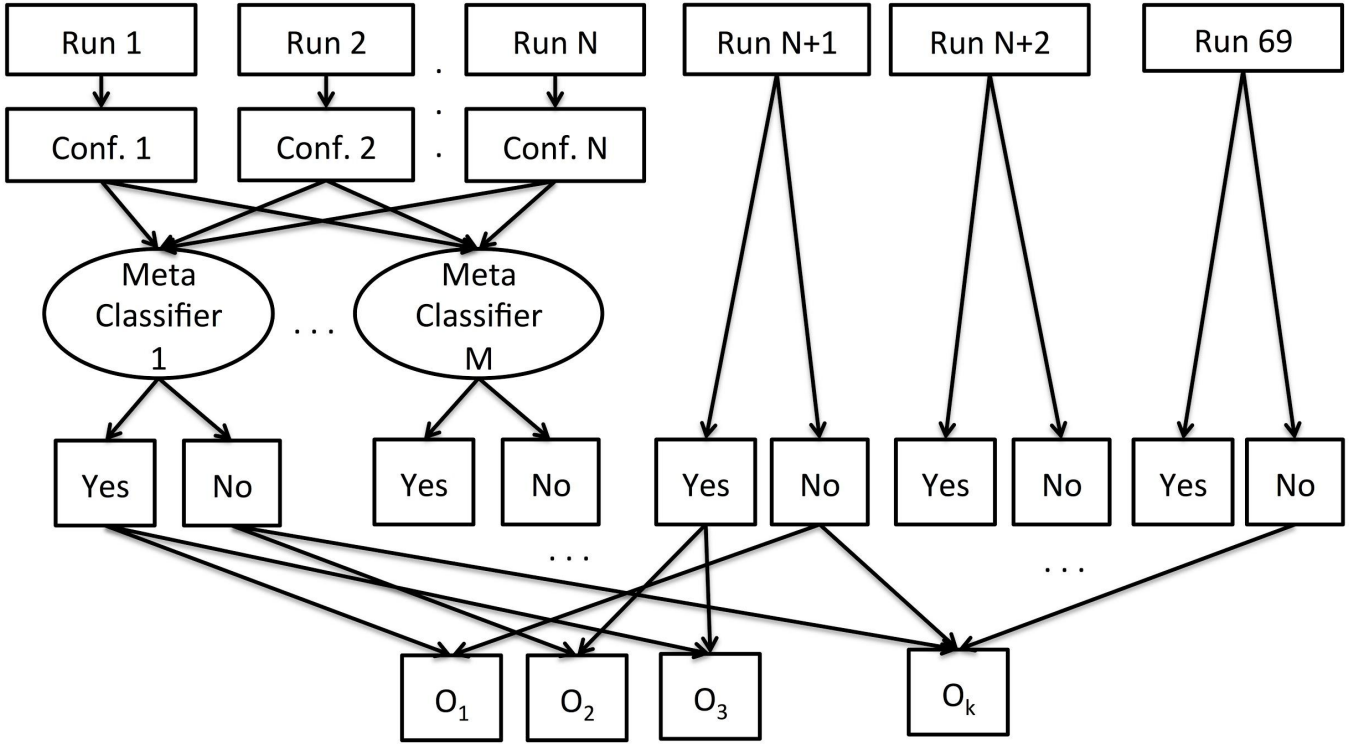
Consensus Maximization Fusion



Add runs that can not fit into the stacked ensemble method. We treat these runs as 2-Class Clusters

DSR @ UF
Data Science Research

Consensus Maximization Fusion

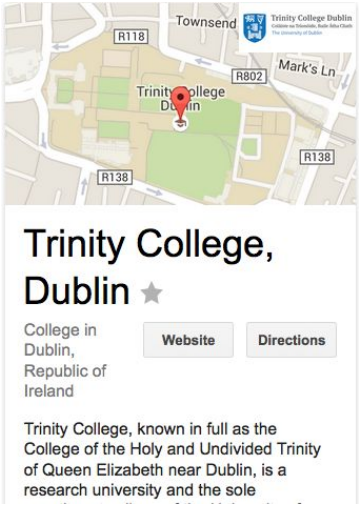


Jing Gao, Feng Liang, Wei Fan, Yizhou Sun, and Jiawe Han. 2009. Graph-based consensus maximization among multiple supervised and unsupervised models. In Advances in Neural Information Processing Systems, pp 585–593.

DSR @ UF
Data Science Research

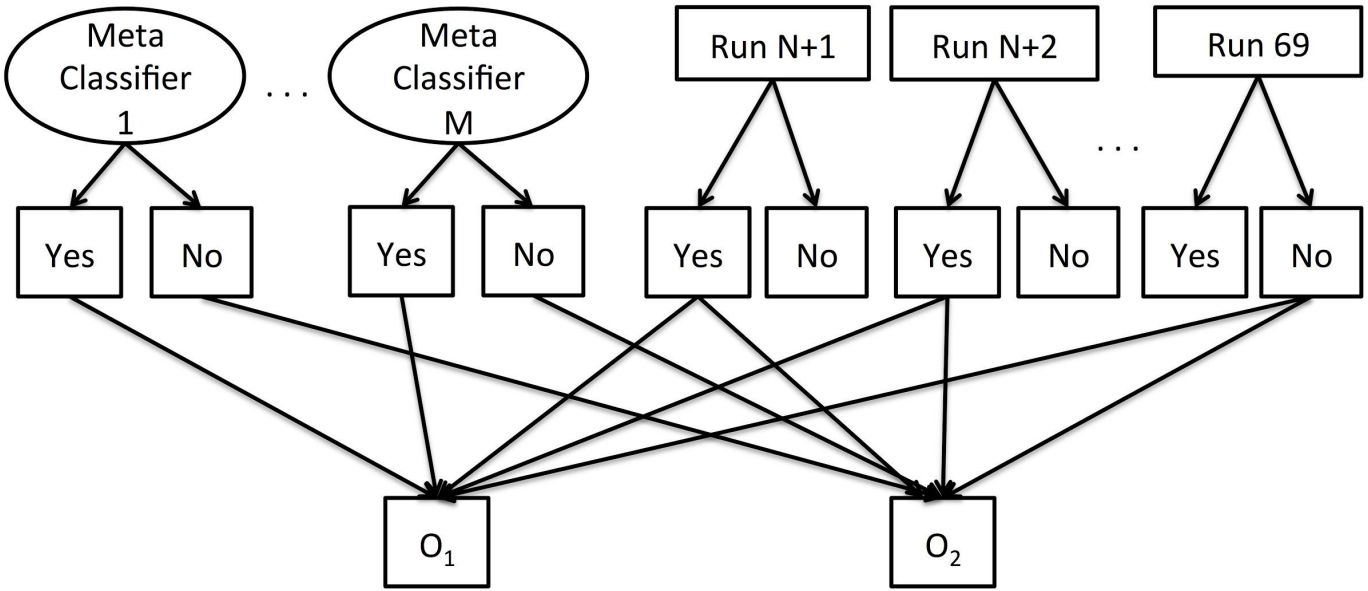
Consensus Max. Fusion - Example

- Consider the following queries
  - O1 = (Marion Hammer, per:title, president)
  - O2 = (Dublin, gpe:headquarters_in_city,trinity college)



DSR @ UF
Data Science Research

Consensus Max. Fusion - Example



Meta-Classifiers: 6 Yes – 0 No

Clusters: 46 Yes - 16No

Meta-Classifiers: 0 Yes – 6 No

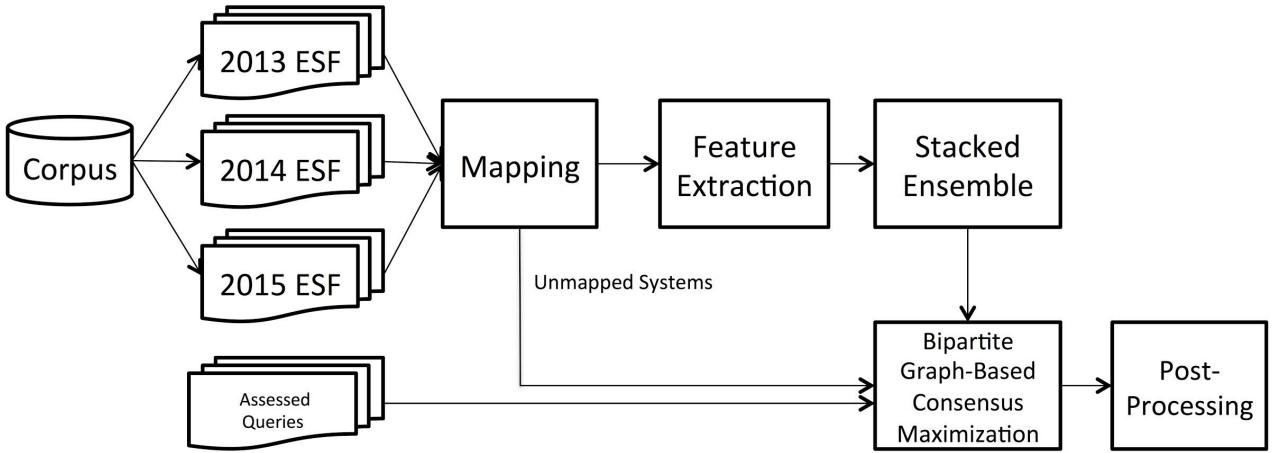
Clusters: 34 Yes - No 28

DSR @ UF
Data Science Research

Consensus Max. Fusion

- Combine outputs of multiple supervised and unsupervised models for better classification.
- The predicted labels should agree with the base supervised models but adds unsupervised evidence.
- Model combination at output level is needed in KBP applications where there is no access to individual extractors.

Consensus Maximization Fusion Pipeline



DSR @ UF
Data Science Research

Mapping

- Runs from teams that participated in previous years are mapped together and ranked using the corresponding assessments.
- 2015 runs, are ranked based on the small assessment file provided for the task.
- The best run of each mapped team is then passes to the feature extraction module.
- All other runs are passed directly to BGCM.

Feature Extraction

- Same as the SFV Stack Ensemble System
  - Probabilities
  - Relation
  - Provenance

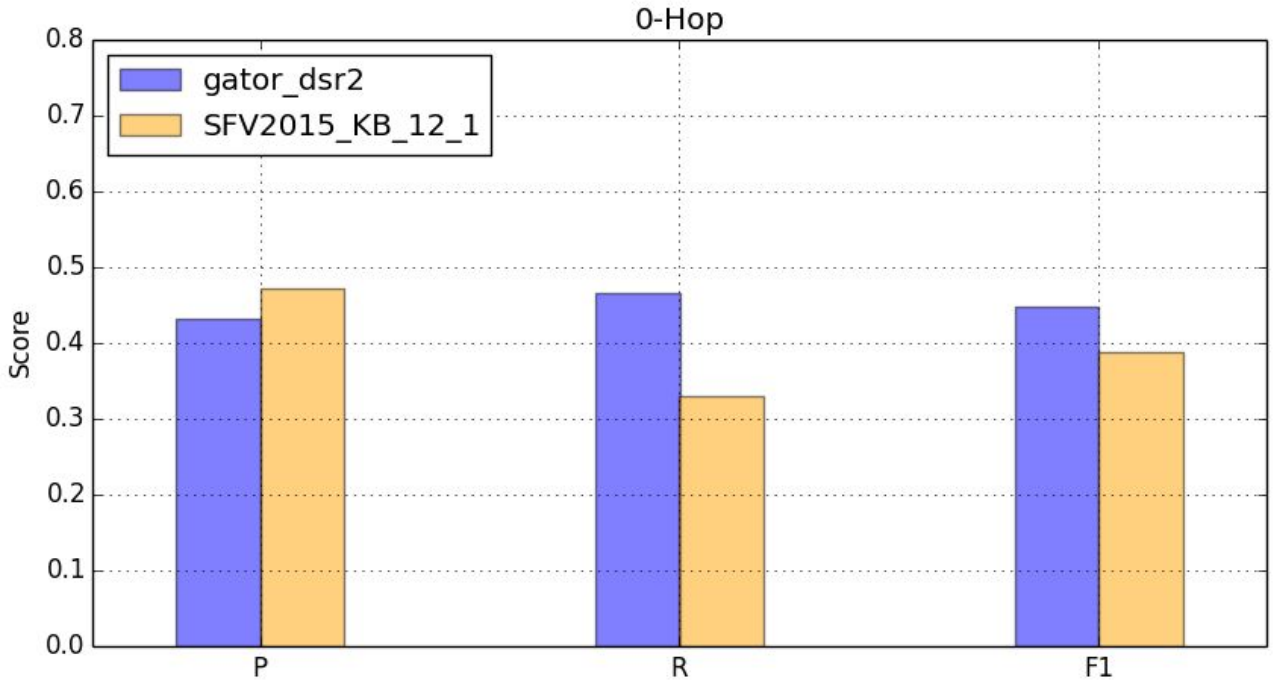
Post-processing

- Filter ensemble of all 0–hop queries
  - Enforce single-values relations by selecting the one with highest probability
  - For every slot filler classified as true, select the provenance of the slot filler with highest probability.
- For every 1-hop query in the ensemble
  - Enforce its 0-hop result is in the ensemble

Submitted Runs

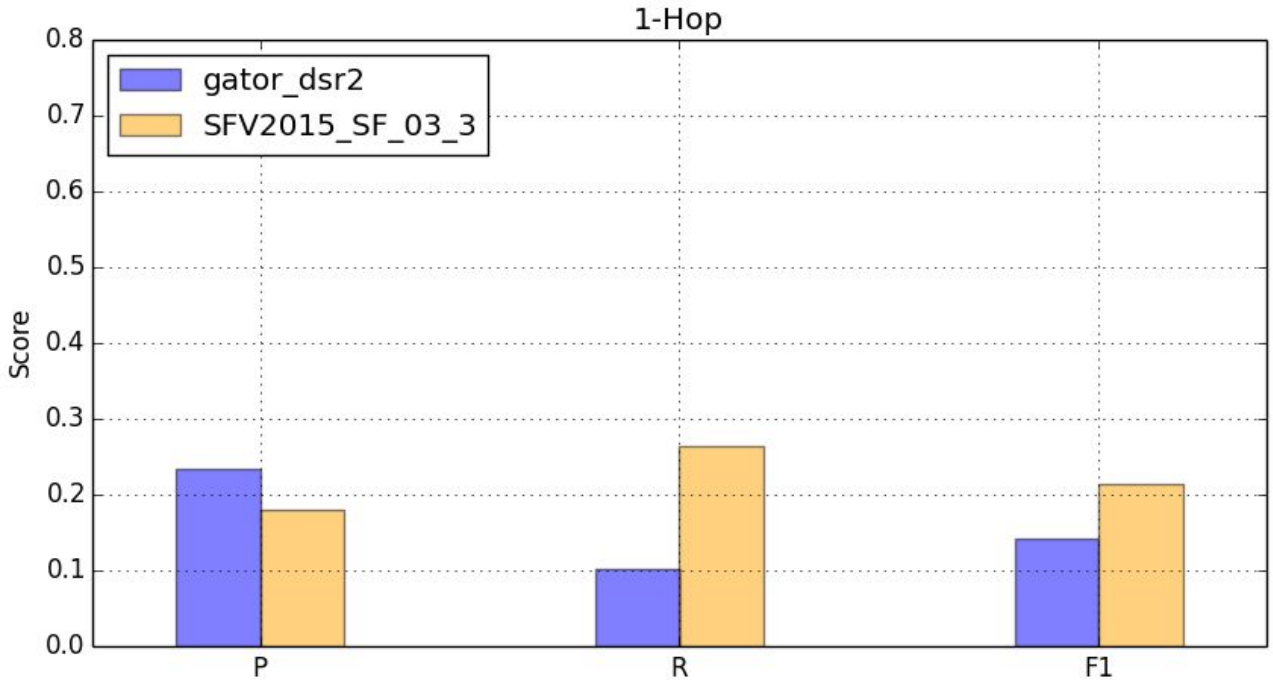
- 2013-2014: Run 1
  - Meta-classifiers trained with samples from 7 teams.
  - BGCM: 6 meta-classifiers and 62 runs
- 2014: Run 2
  - Meta-classifiers trained with samples from 12 teams.
  - BGCM: 6 meta-classifiers and 57 runs
- Run 3
  - Use all meta classifiers from Runs 1 and 2
  - BGCM: 12 meta-classifiers and 57 runs

Results - 2015 CSSF



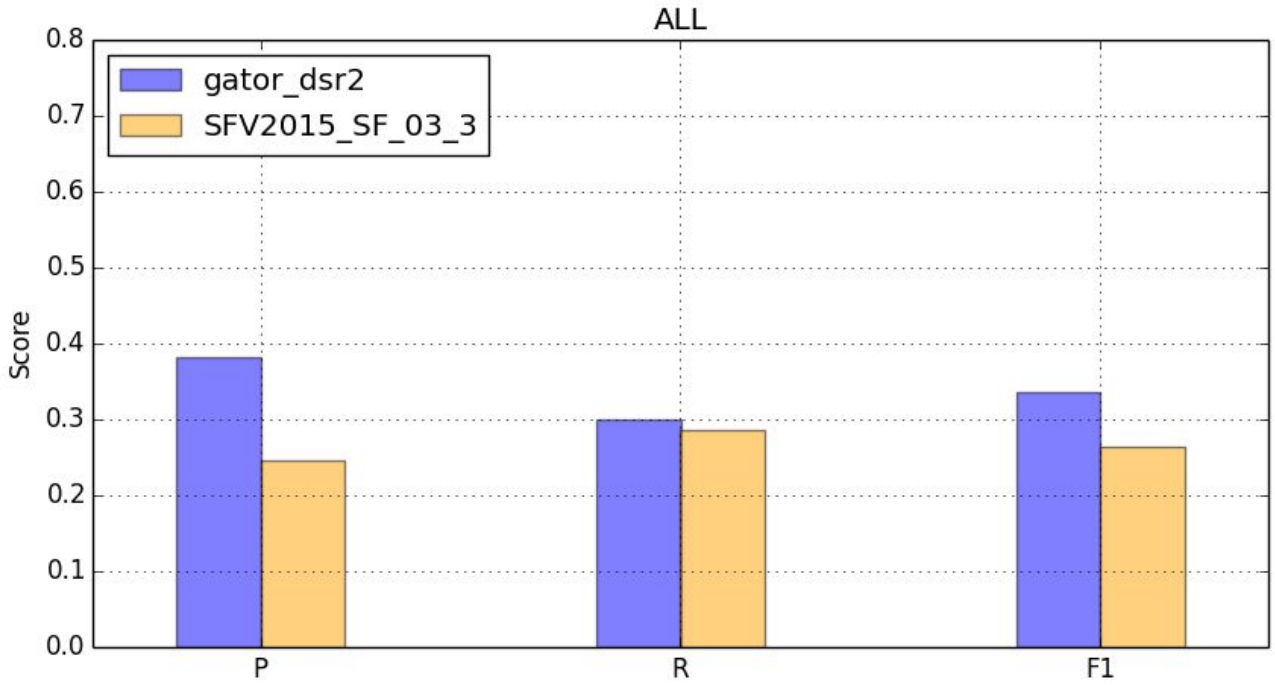
DSR @ UF
Data Science Research

Results - 2015 CSSF



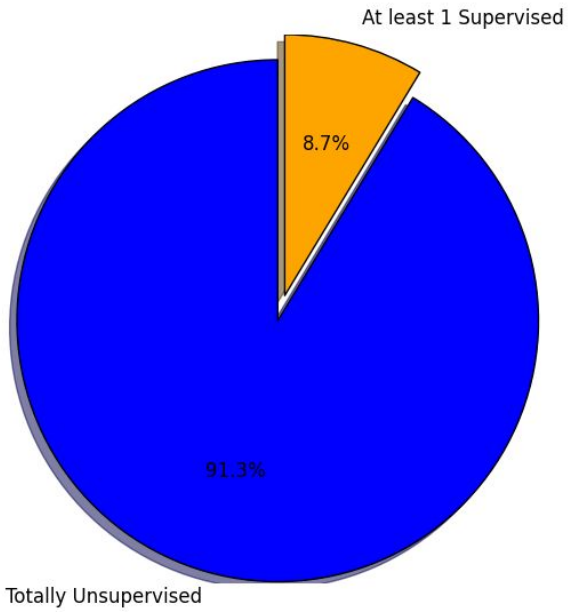
DSR @ UF
Data Science Research

Results - 2015 CSSF



DSR @ UF
Data Science Research

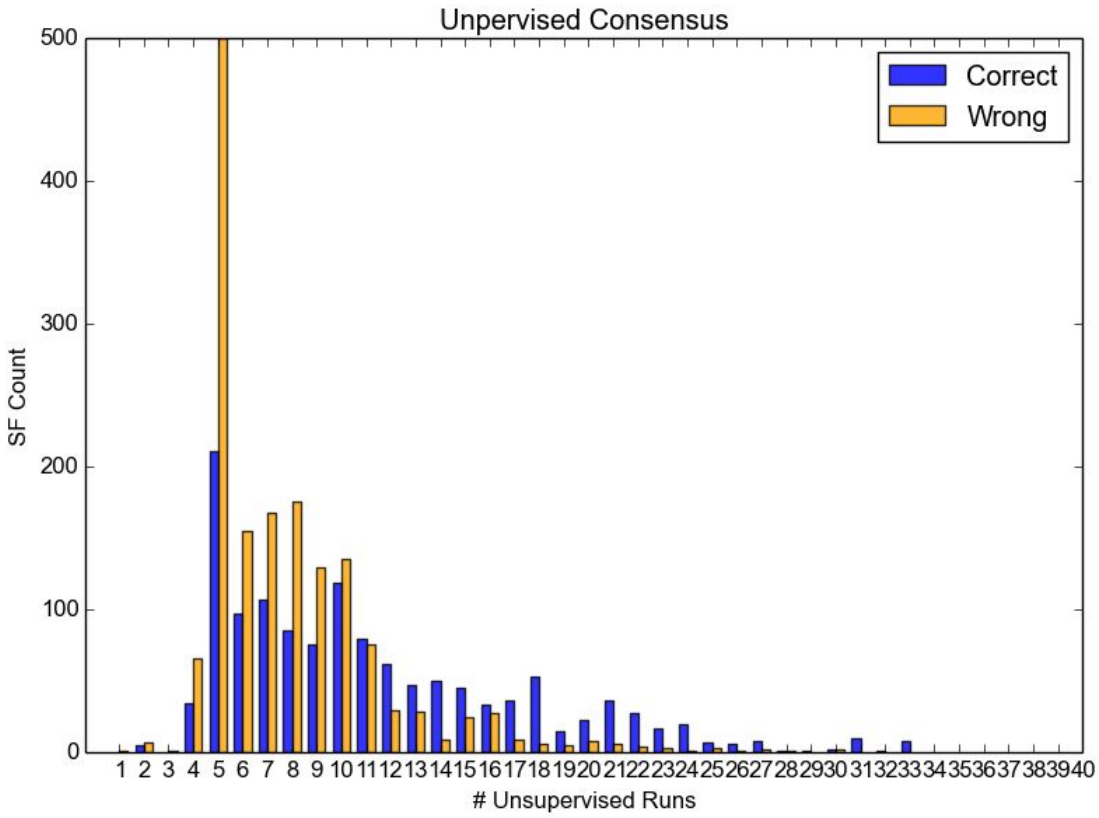
Analysis Run 2



The majority of the slot fillers included in our best run come from unsupervised consensus

DSR @ UF
Data Science Research

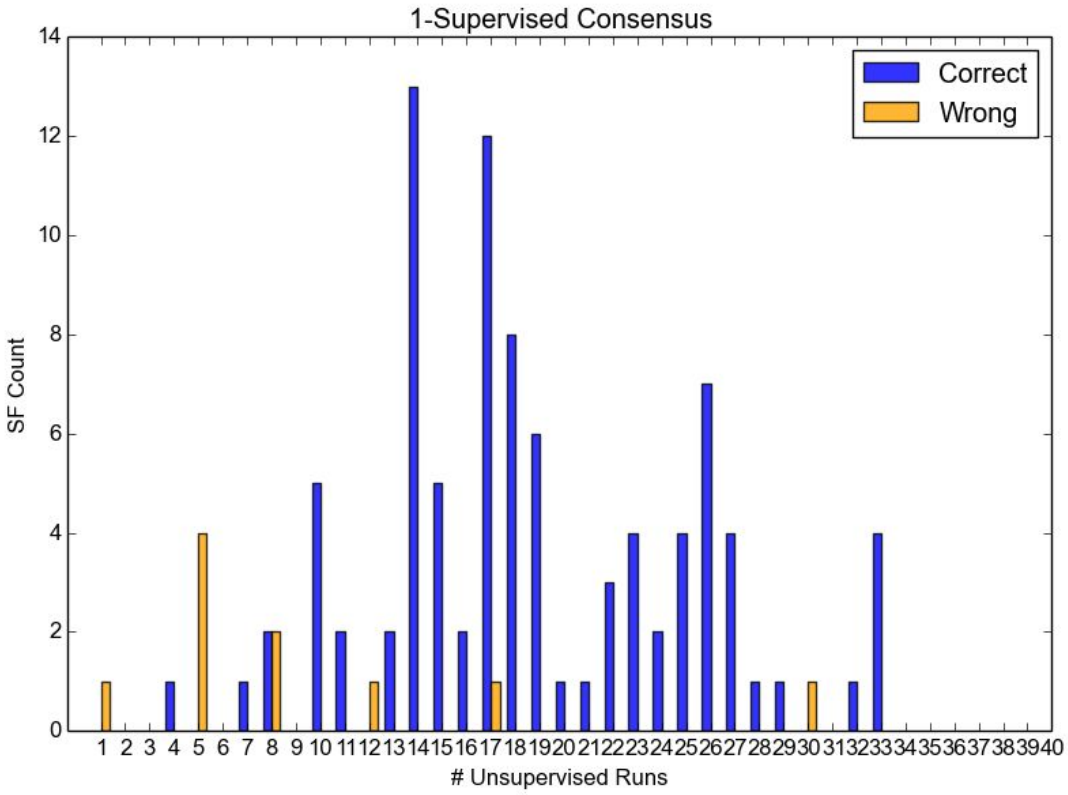
Analysis Run 2



- Answers come from unsupervised consensus
  - All supervised outputs classified them as negative
  - Not enough evidence
- As more unsupervised runs reach consensus, there are more correct than incorrect fillers.
- The Recall of the system is improved

DSR @ UF
Data Science Research

Analysis Run 2



- At least one stacked ensemble model classified as positive.
- Supervised evidence helps improve precision.
- The higher the consensus with the unsupervised clusters the system filters better.

DSR @ UF
Data Science Research

Questions?