Stacked Ensembles
of Information Extractors for
Knowledge-Base Population

Nazneen Rajani
Raymond J. Mooney,
Vidhoon Vishwanathan and Yinon Bentor
University of Texas at Austin

Knowledge-Base Population (KBP)

- Annual evaluation of relation extraction from natural language documents organized by NIST.

- English Slot Filling (ESF) task:

per: Barack Obama
country_of_birth United States
spouse Michelle Obama
children Malia Obama Sasha Obama

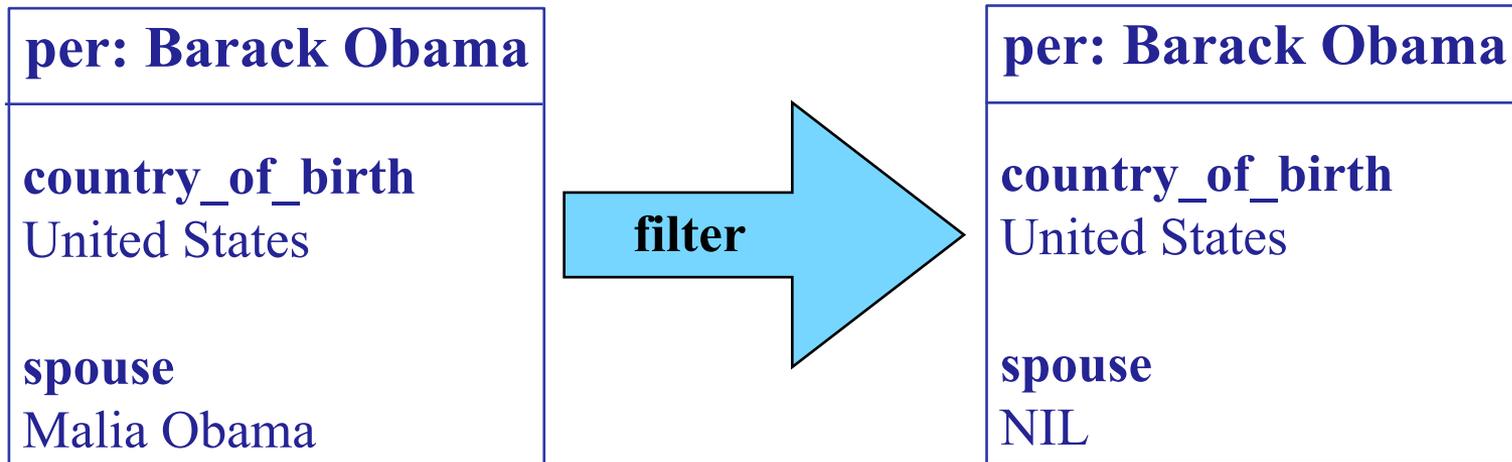| org: Microsoft |
| --- |
| **city_of_headquarters**
Redmond |
| **website**
microsoft.com |
| **subsidiaries**
Skype
Nokia |

KBP Provenance

- System's must provide information on where the evidence for each slot fill is in the document corpus.
- Given by:
  - Doc ID
  - Start Offset
  - End Offset

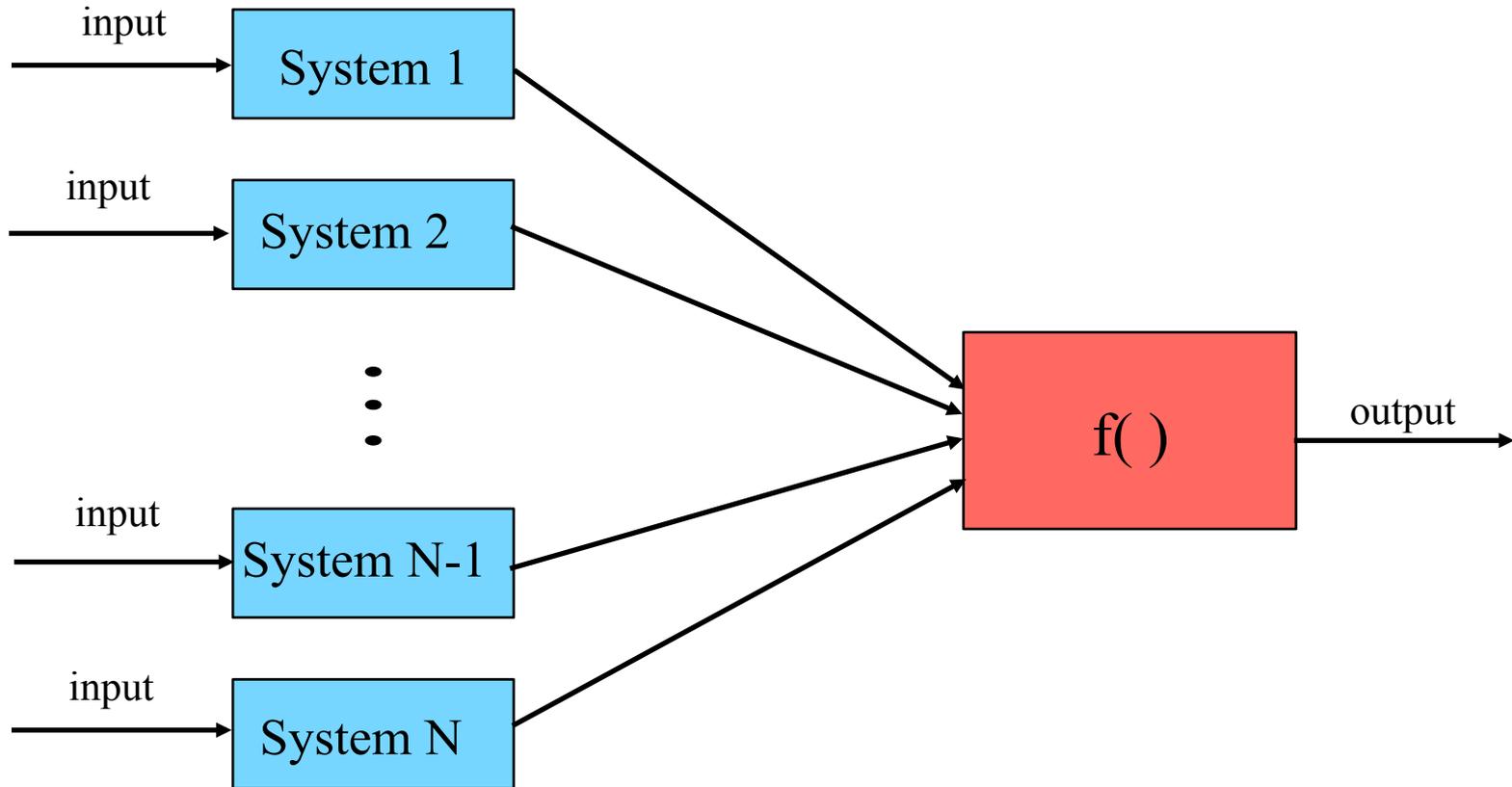| org: Microsoft |
|---|
| \<eng-NG-31-1007\> : Microsoft is a technology company headquartered in Redmond, Washington, that develops … |
| **city_of_headquarters** Redmond **Doc ID** eng-NG-31-1007 **Start Offset** 48 **End Offset** 54 |

KBP Slot Filler Validation

- Aim: Improve precision of individual systems.

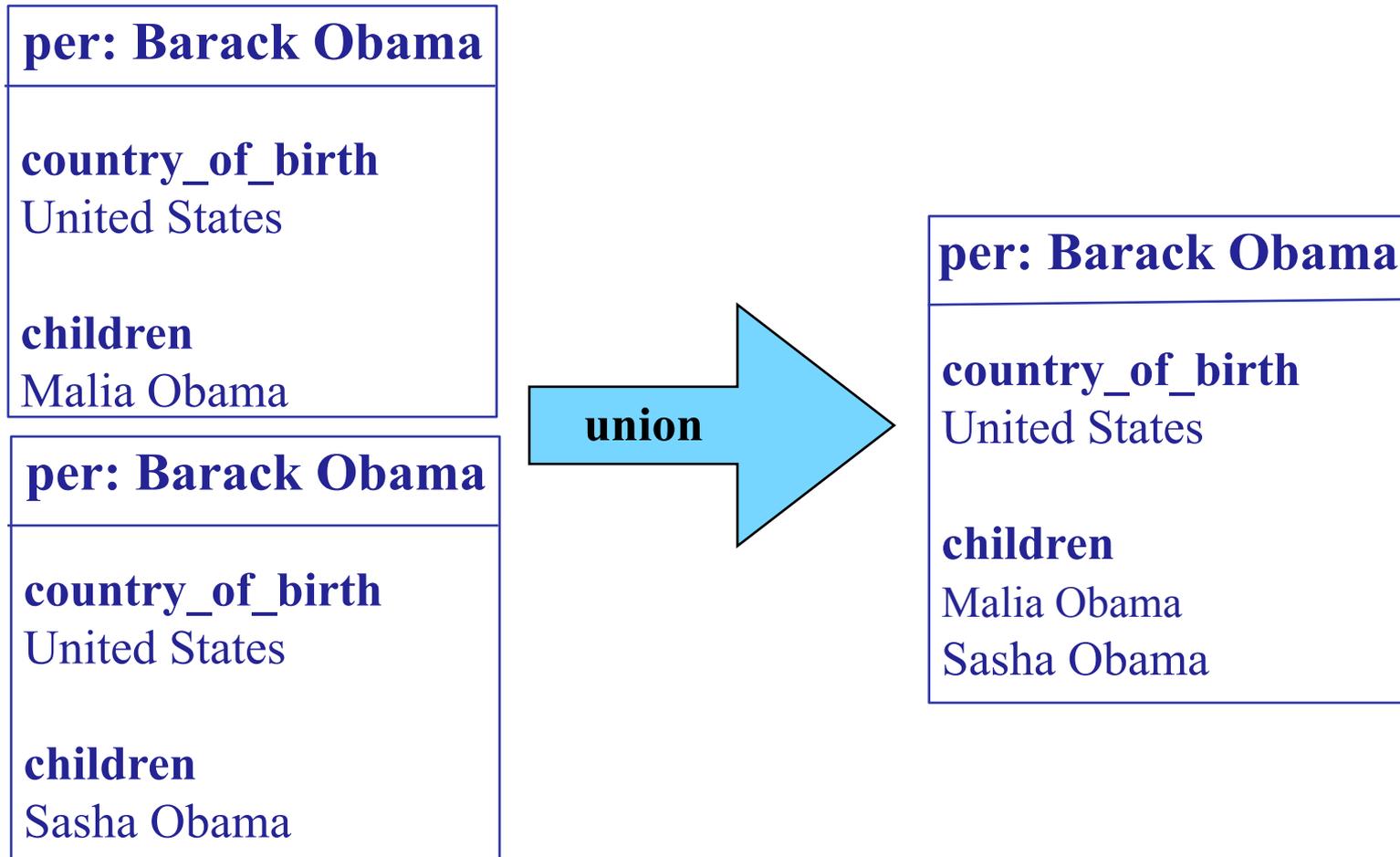- Input is system outputs from the ESF task.

- Output is filtered slot fills.

per: Barack Obama		per: Barack Obama
country_of_birth United States **spouse** Malia Obama	**filter** →	**country_of_birth** United States **spouse** NIL

Ensembling

KBP Slot Filler Validation

- Ensembling used to improve recall as well

per: Barack Obama

country_of_birth
United States

children
Malia Obama

per: Barack Obama

country_of_birth
United States

children
Sasha Obama

union

per: Barack Obama

country_of_birth
United States

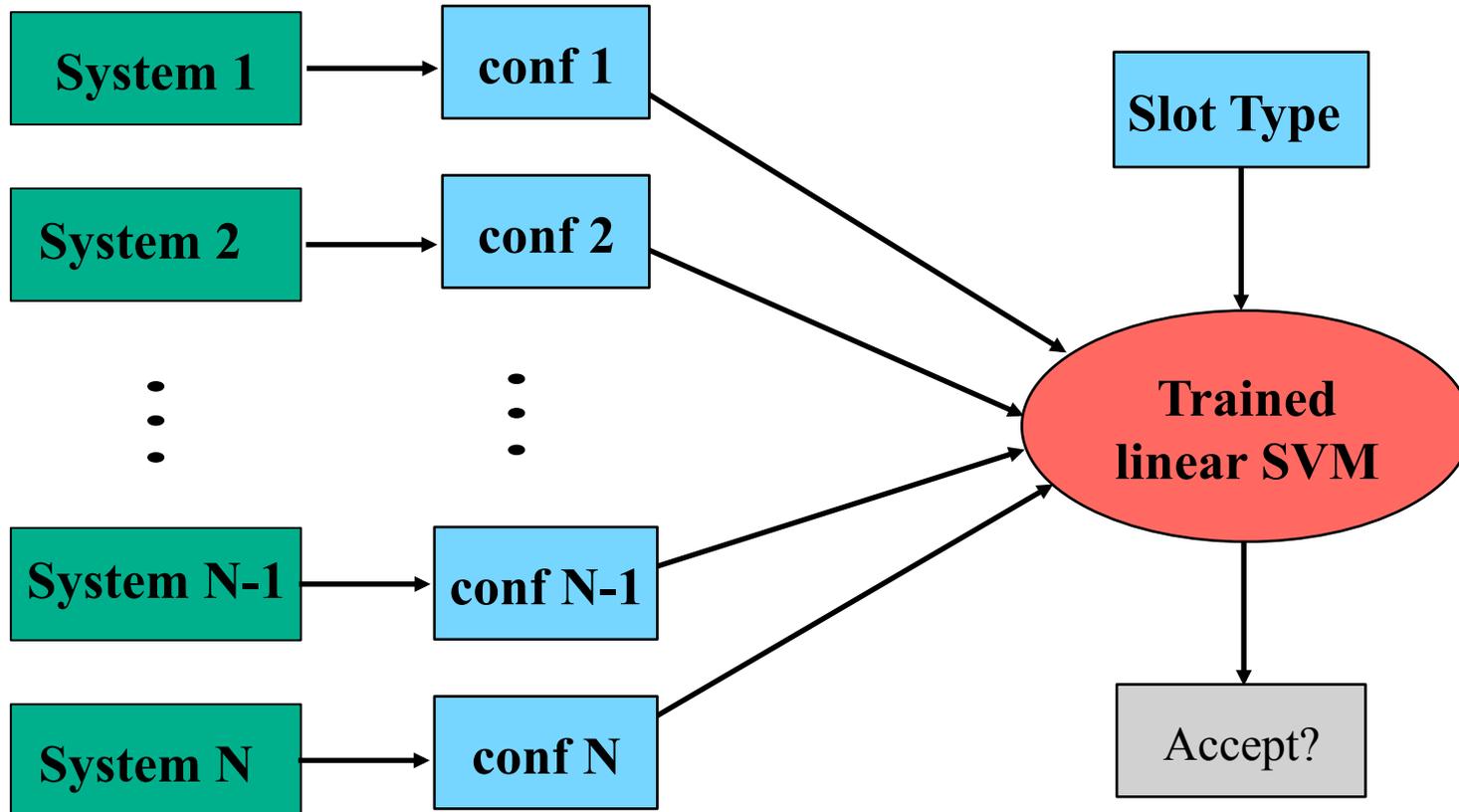**children**
Malia Obama
Sasha Obama

Stacking

(Wolpert, 1992)

For a given proposed slot-fill, e.g. `spouse(Barack, Michelle)`, combine confidences from multiple systems:
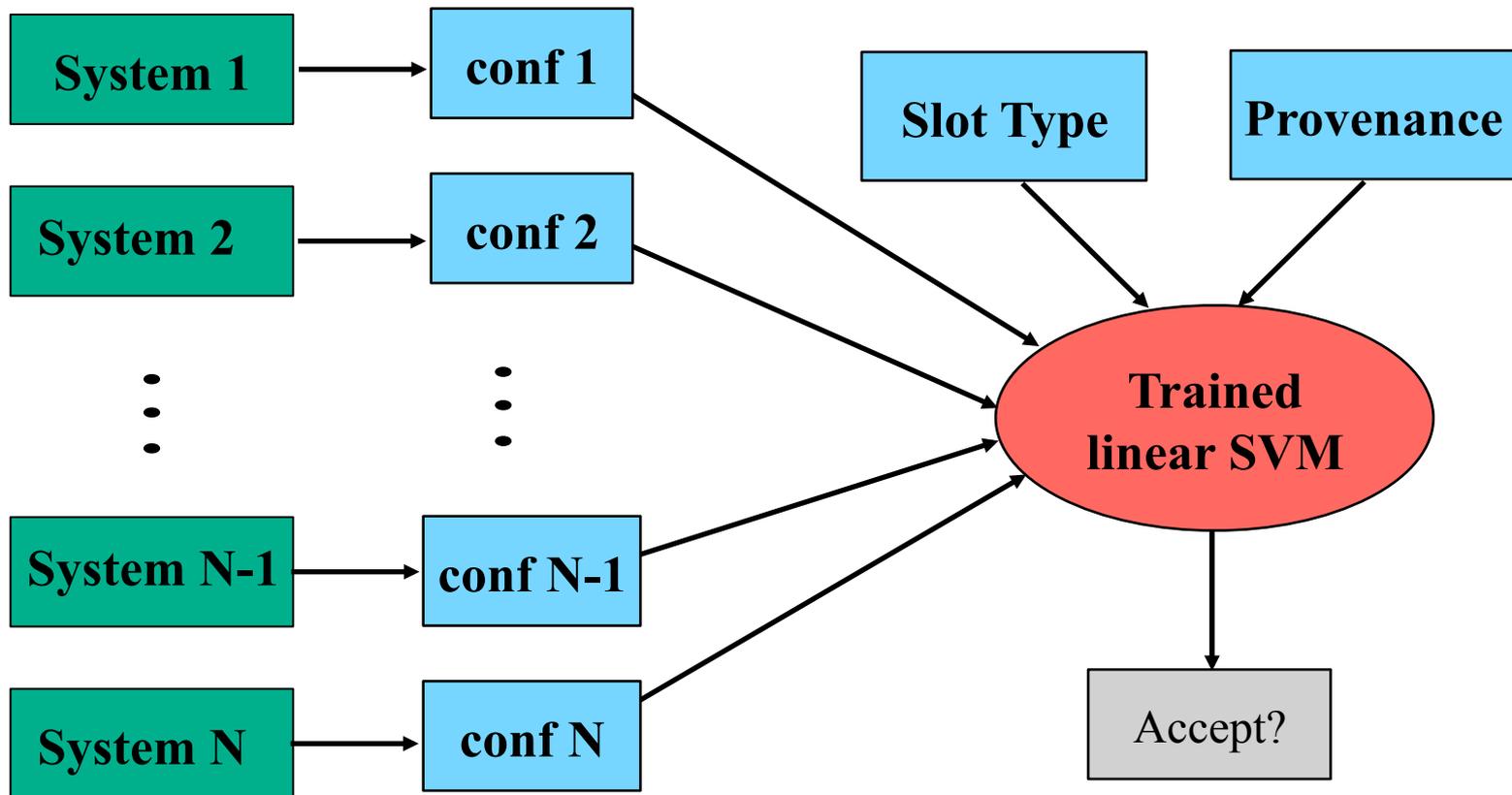
Stacking with Features

For a given proposed slot-fill, e.g. `spouse(Barack, Michelle)`, combine confidences from multiple systems:

Stacking with Features

For a given proposed slot-fill, e.g. `spouse(Barack, Michelle)`, combine confidences from multiple systems:

| System 1 | → | conf 1 |
| System 2 | → | conf 2 |

Slot Type   Provenance

Trained linear SVM

System N-1 → conf N-1

System N → conf N

Accept?
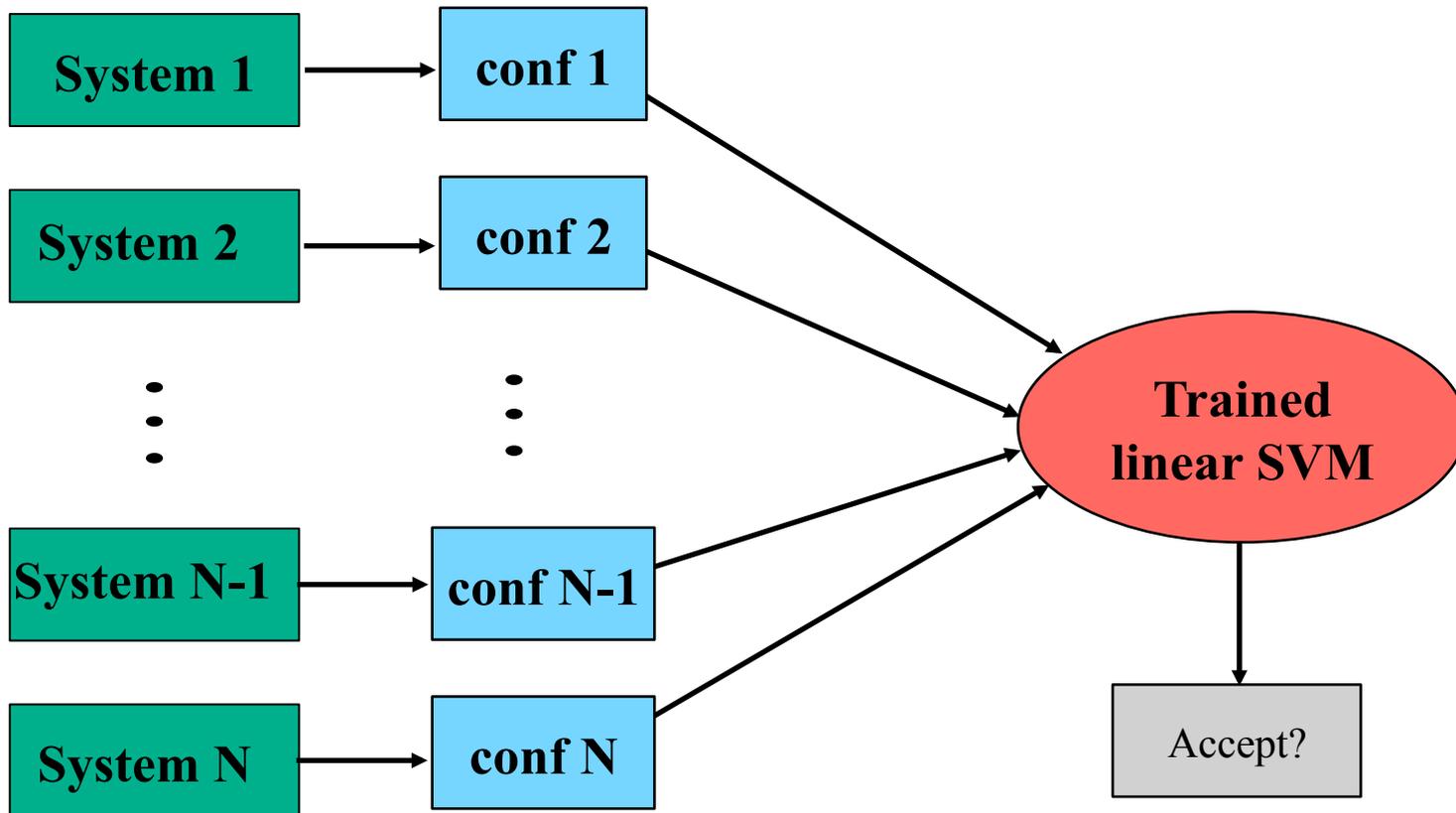
Document Provenance Feature

- For a given query and slot, for each system, *i,* there is a feature DP_i:

  – *N* systems provide a fill for the slot.

  – Of these, *n* give same provenance *docid* as *i.*

  – $DP_i = n/N$ is the document provenance score.

- Measures extent to which systems agree on document provenance of the slot fill.

Offset Provenance Feature

- Degree of overlap between systems' provenance strings (prov).
- Uses Jaccard similarity coefficient.
- For a given query and slot, for each system, *i,* there is a feature OP_i :
  - *N* systems provide a fill with same *docid*
  - Offset provenance for a system *i* is calculated as:

$$OP_i = \frac{1}{|N|} \times \sum_{j \in N, j \neq i} \frac{|\mathsf{prov(i)} \cap \mathsf{prov(j)}|}{|\mathsf{prov(i)} \cup \mathsf{prov(j)}|}$$

  - Systems with different *docid* have zero OP

Offset Provenance Score

Offsets	System 1	System 2	System 3
Start Offset	1	4	5
End Offset	9	7	12
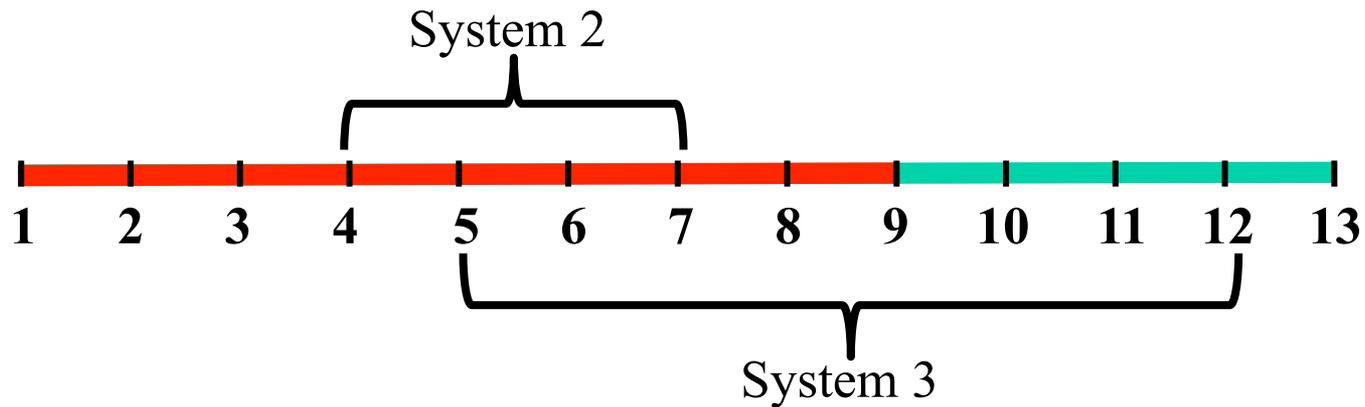
System 2

System 3

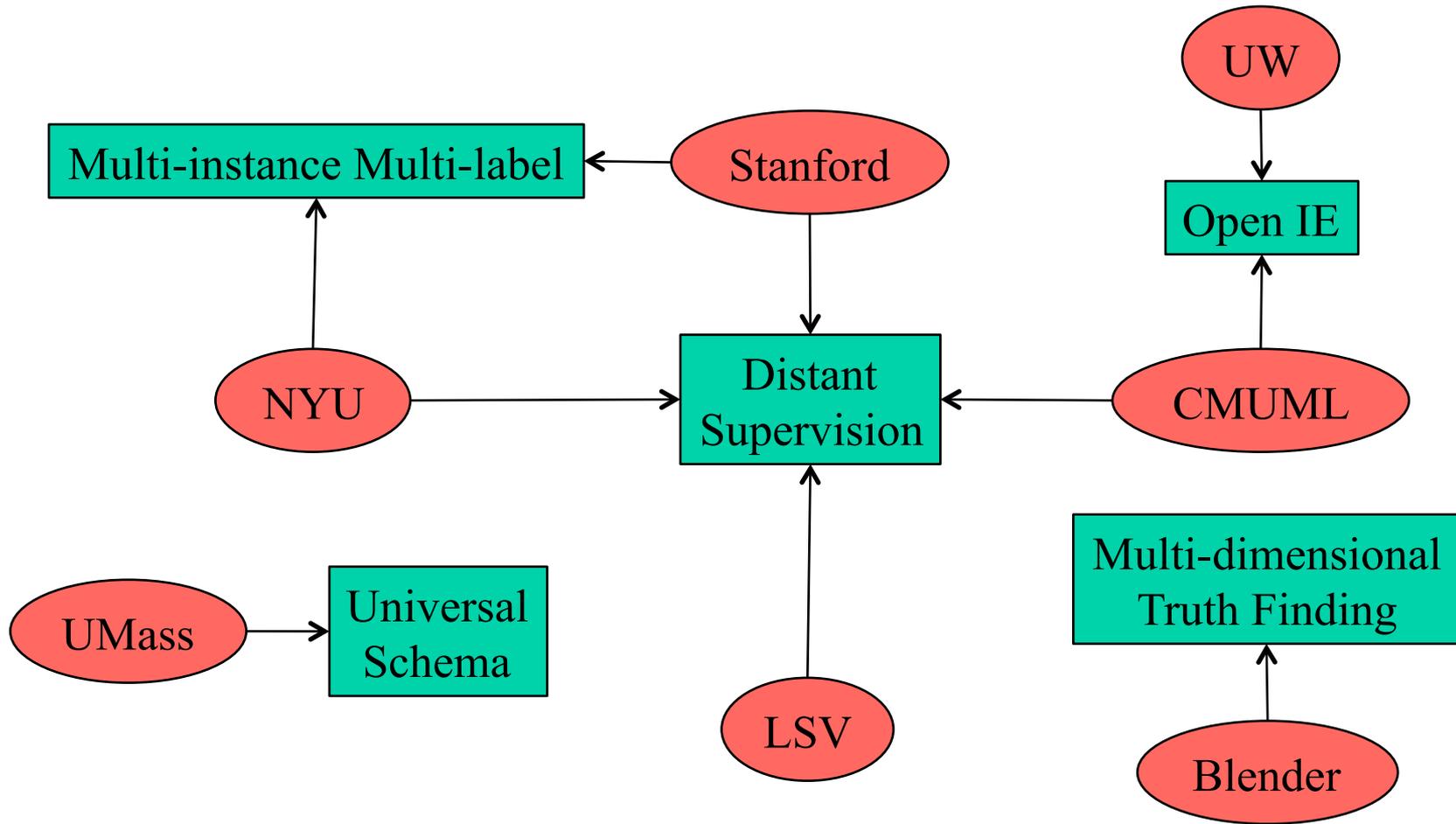$$OP_1 = \frac{1}{2} \times \left(\frac{4}{9} + \frac{5}{12} \right)$$

Datasets

- 2014 Slot Filler Validation (SFV) data
  - 17 teams
  - 65 systems
- Ten Common Systems that participated both in 2013 and 2014 English Slot Filling (ESF) task:
  - LSV
  - IIRG
  - UMASS_IESL
  - Stanford
  - BUPT_PRIS
  - RPI_BLENDER
  - CMUML
  - NYU
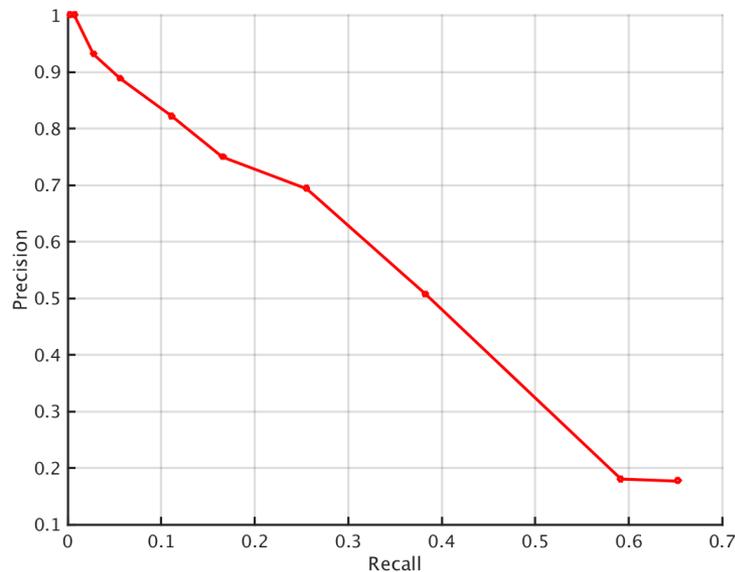 - Compreno
  - UWashington

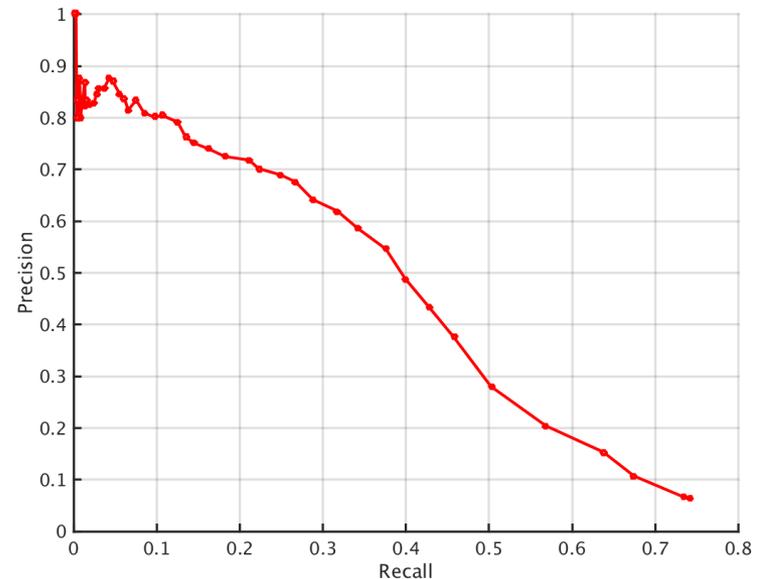# Common Systems Overview

Baselines

- Union

  - Combine systems for maximizing recall

  - List valued slot fills => always included

  - Single valued slot fills => highest confidence

Baselines

- Voting
  - Combine systems for maximizing F1
  - Vary threshold on #systems that must agree
  - Learn threshold on 2013 data

Common Systems Dataset (3)

SFV Dataset (10)

KBP English Slot Filling (ESF) Results

2014 Slot Filler Validation (SFV) Data

Baseline	Precision	Recall	F1
Union	0.067	**0.762**	0.122
Voting	**0.641**	0.288	**0.397**

Common systems for 2013 and 2014 ESF task

Approach	Precision	Recall	F1
Union	0.176	**0.647**	0.277
Voting	**0.694**	0.256	0.374
Best ESF system in 2014 (Stanford)	0.585	0.298	0.395
Stacking	0.606	0.402	0.483
Stacking + Relation	0.607	0.406	0.486
Stacking + Provenance + Relation	0.541	0.466	**0.501**

Slot Filler Validation (SFV) Key

- Assessment key for ESF has:
  - Human annotated slot fills
  - Pooled slot fills of systems
- Slot filler validation key used by Sammons et al. (2014):
  - Only includes the correct pooled slot fills
  - We use this key to compare our approach to theirs

KBP Slot Filler Validation Results

2014 Slot Filler Validation (SFV) Data

Baseline	Precision	Recall	F1
Union	0.054	**0.877**	0.101
Voting	**0.637**	0.406	**0.496**

Common systems for 2013 and 2014 ESF task

| Approach	Precision	Recall	F1
Union	0.177	**0.922**	0.296
Voting	**0.694**	0.256	0.374
Best SFV system in 2014 (UIUC)	0.457	0.507	0.481
Stacking	0.613	0.562	0.586
Stacking + Relation	0.613	0.567	0.589
Stacking + Provenance + Relation	0.659	0.560	**0.606**

Learning Curve

- Systems change each year.
- Still useful to train on past data.
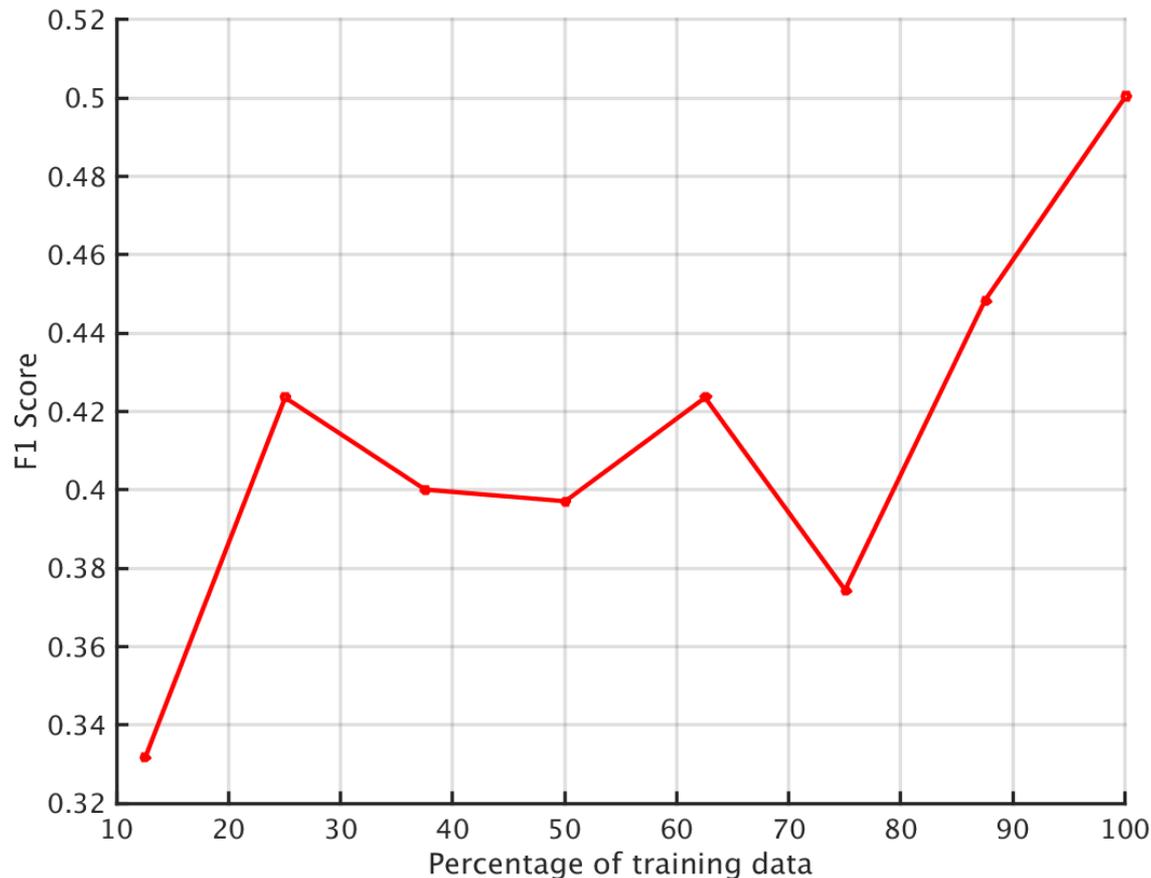
Incremental Training on Systems

- Sort the common systems based on their performance.
- Train the classifier adding one system at each step.
- Test on 2014 data.
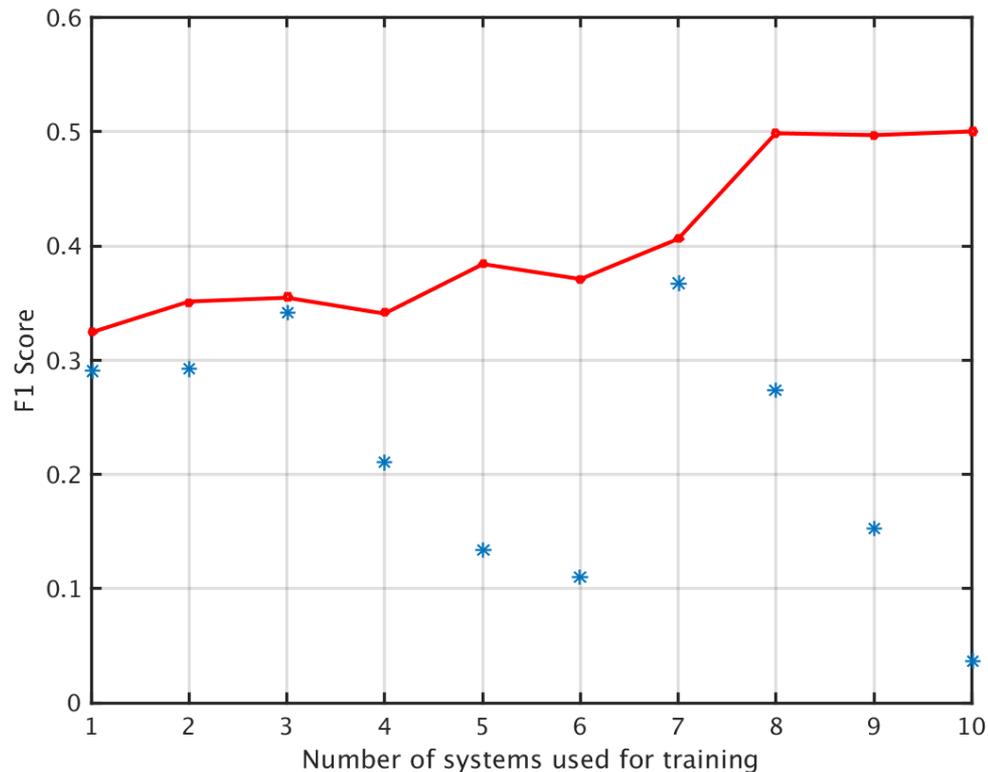
Conclusion

- Stacked meta-classifier beats the best performing 2014 KBP ESF system by an F1 gain of **11** points.

- Features that utilize provenance information improve stacking performance.

- Ensembling has clear advantages but naive approaches such as voting do not perform as well.

- Although systems change every year, there are advantages in training on past data.

Questions?