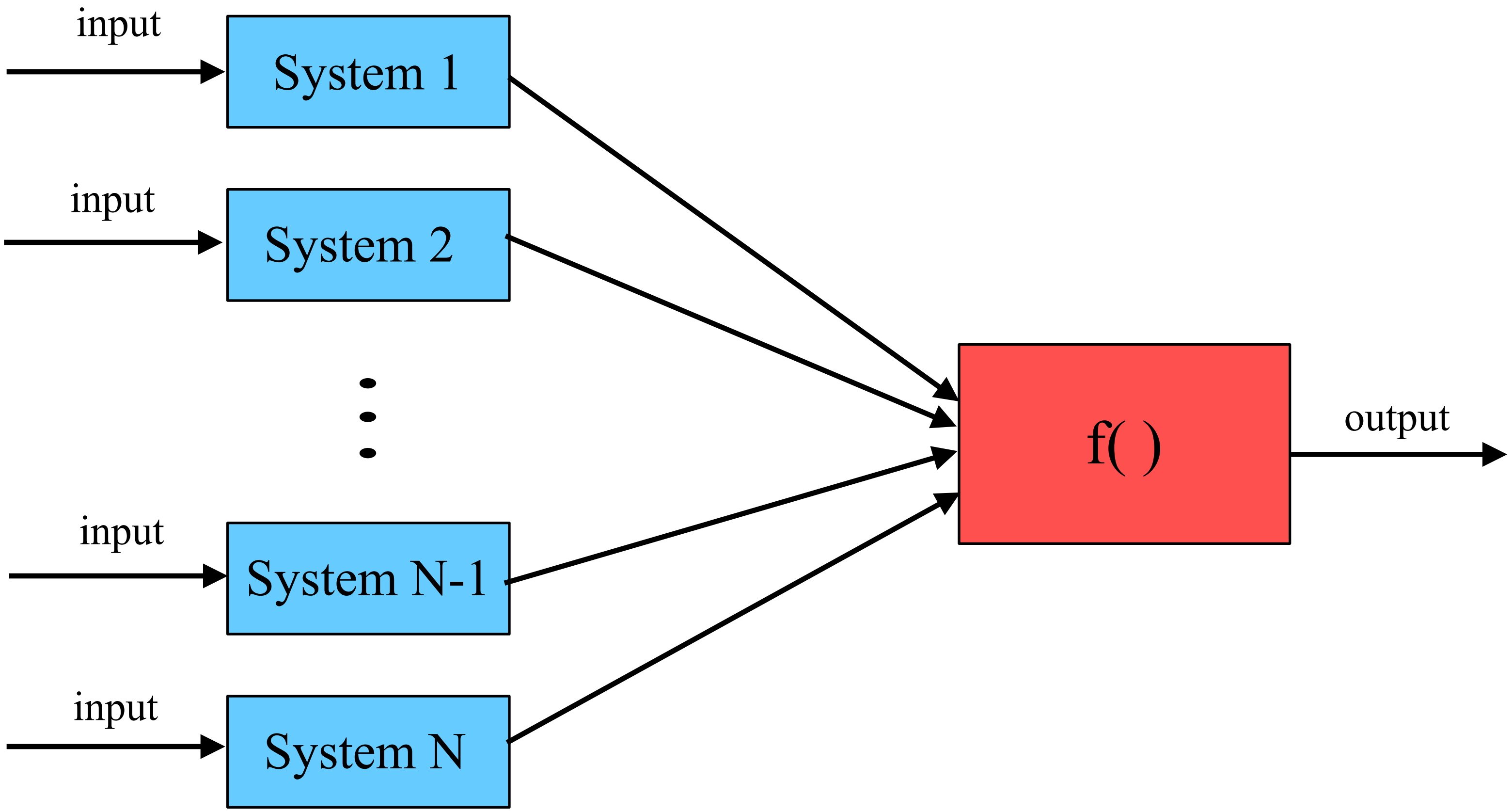
Stacked Ensembles of Information Extractors for Knowledge-Base Population

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

Ensembling


input
System 1
input
System 2
input
System N-1
input
System N
f( )
output

Background

- English Slot Filling(ESF) task
  - 52 submissions in 2013
  - 65 submissions in 2014

per: Barack Obama

country_of_birth
United States

spouse
Michelle Obama

children
Malia Obama
Sasha Obama

org: Microsoft

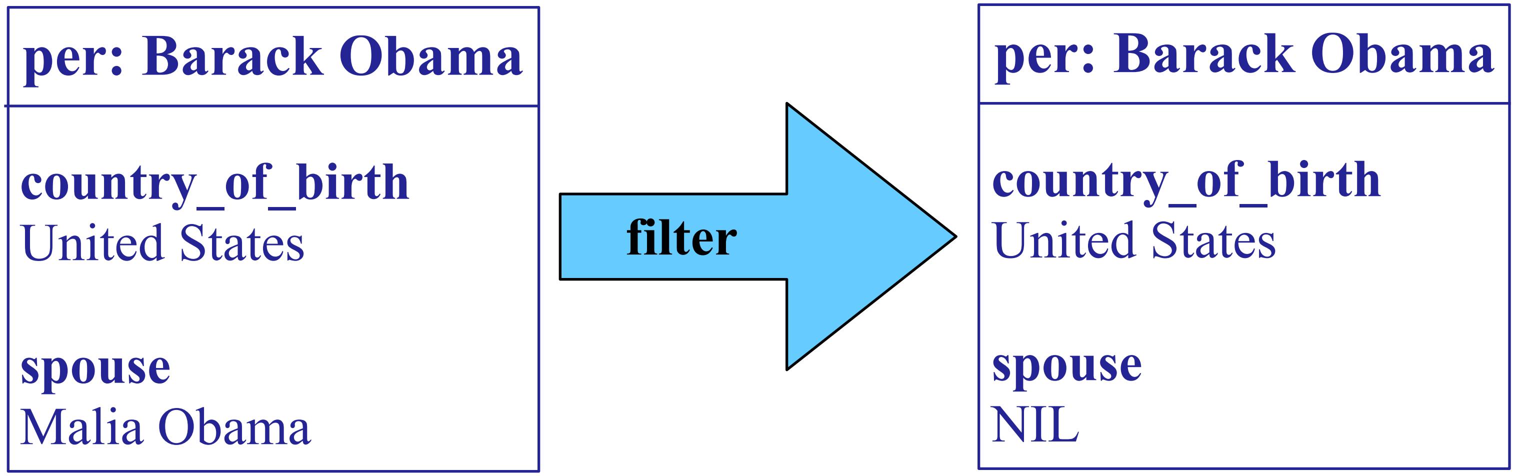
city_of_headquarters
Redmond

website
microsoft.com

subsidiaries
Skype
Nokia

Background

- Slot Filler Validation (SFV) task
  - Improving precision of individual systems
  - Input is outputs from the ESF task
  - Output is filtered slot fills
  - Can be used for improving recall as well



per: Barack Obama

country_of_birth
United States

spouse
Malia Obama

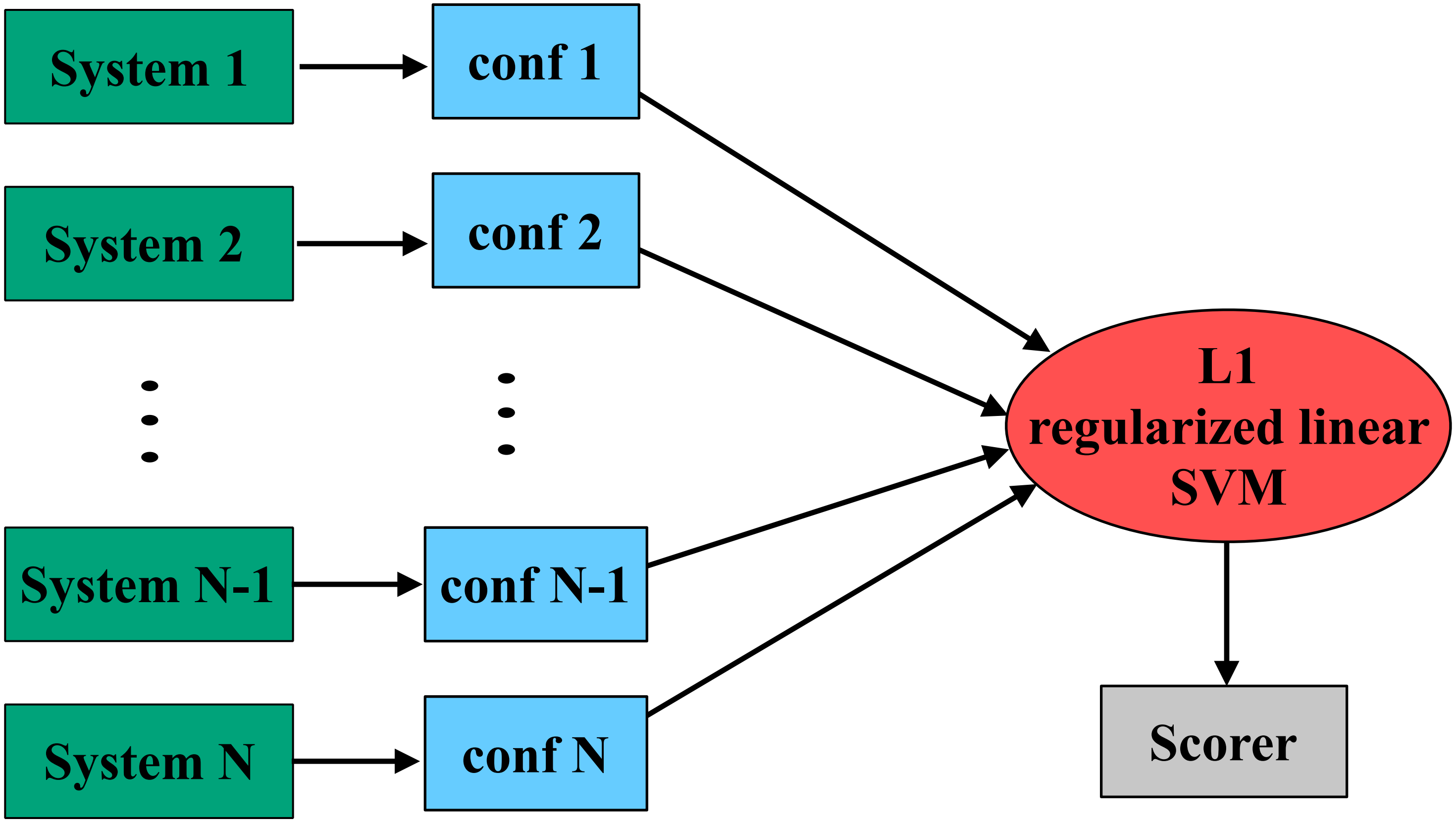
filter

per: Barack Obama

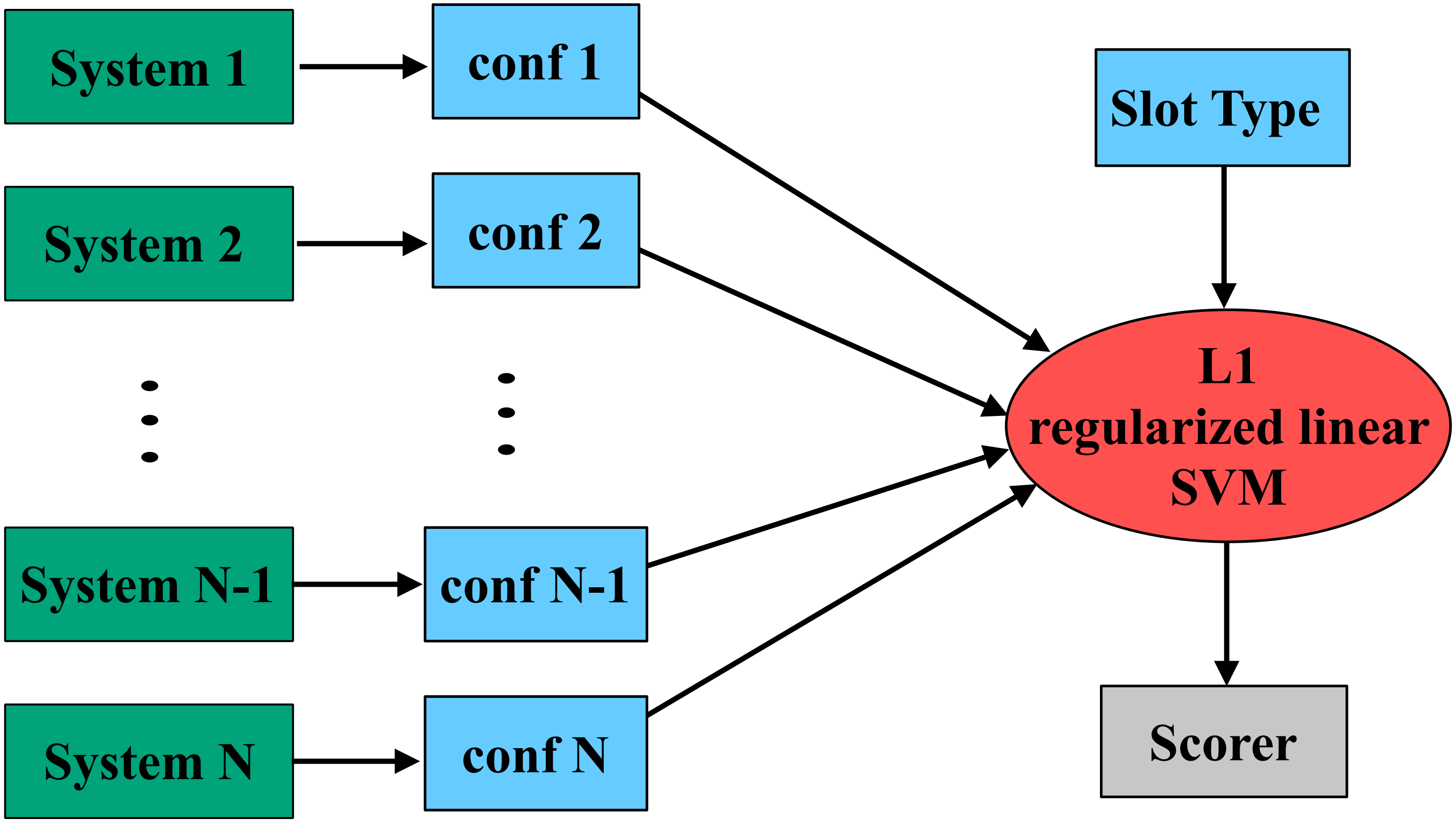
country_of_birth
United States

spouse
NIL

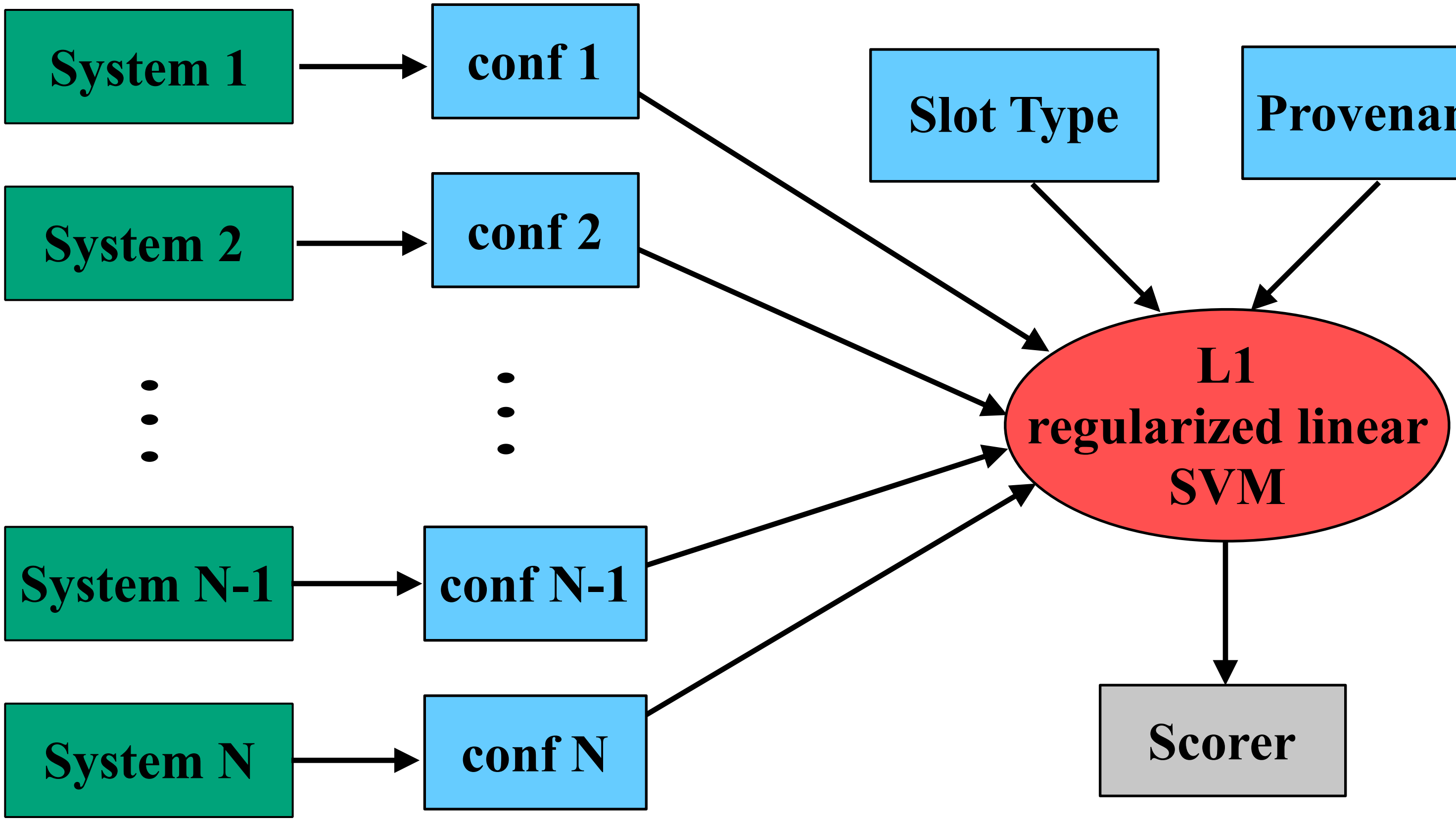
Stacking


System 1
System 2
System N-1
System N
conf 1
conf 2
conf N-1
conf N
L1
regularized linear
SVM
Scorer

Stacking with Features



Stacking with Features



Using Document Provenance

- No access to the corpus
- *docid, startoffset-endoffset*
- Document based provenance score
  - For a given query and slot:
    - *N* systems provide answers
    - *n* give same *docid* as provenance
    - *n/N* is the document provenance score
    - Sums to 1
  - Measures extent to which systems agree on document provenance of the slot fill

Using Offset Provenance

- Offset based provenance (OP) score
  - Degree of overlap between systems' provenance strings (prov)
  - Jaccard similarity coefficient
  - For a given query and slot:
    - *N* systems provide answers with same *docid*
    - OP for a system *j* is calculated as

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

    - Systems with different *docid* have zero OP
    - Used along with document provenance

Datasets

- Ten Common Systems that participated both in 2013 as well as 2014

LSV
IIRG
UMass IESL
Stanford
BUPT PRIS
RPI BLENDER
CMUML
NYU
Compreno
UWashington

- 2014 Slot Filler Validation data
  - 17 teams
  - 65 systems

Baselines

- Union
  - Combine systems for maximizing recall
  - List valued slot fills => always included
  - Single valued slot fills => highest confidence
- Voting
  - Combine systems for maximizing precision
  - Vary threshold on #systems that must agree
  - Learn threshold on 2013 data
  - SFV and common systems datasets

KBP English Slot Filling 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**

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	0.56	**0.606**

Conclusion

- Stacked meta-classifier beats the best performing 2014 ESF system by a F1 score of **11** points
- Provenance without access to corpus is also useful
- Pooling has advantages but naive approaches such as voting do not perform as well