Adding Smarter Systems Instead of Human Annotators: A Case Study in Slot Filling

Zheng Chen, Suzanne Tamang, Adam Lee, Xiang Li, Wen-Pin Lin,
Matthew Snover, Javier Artiles, Marissa Passantino, Heng Ji

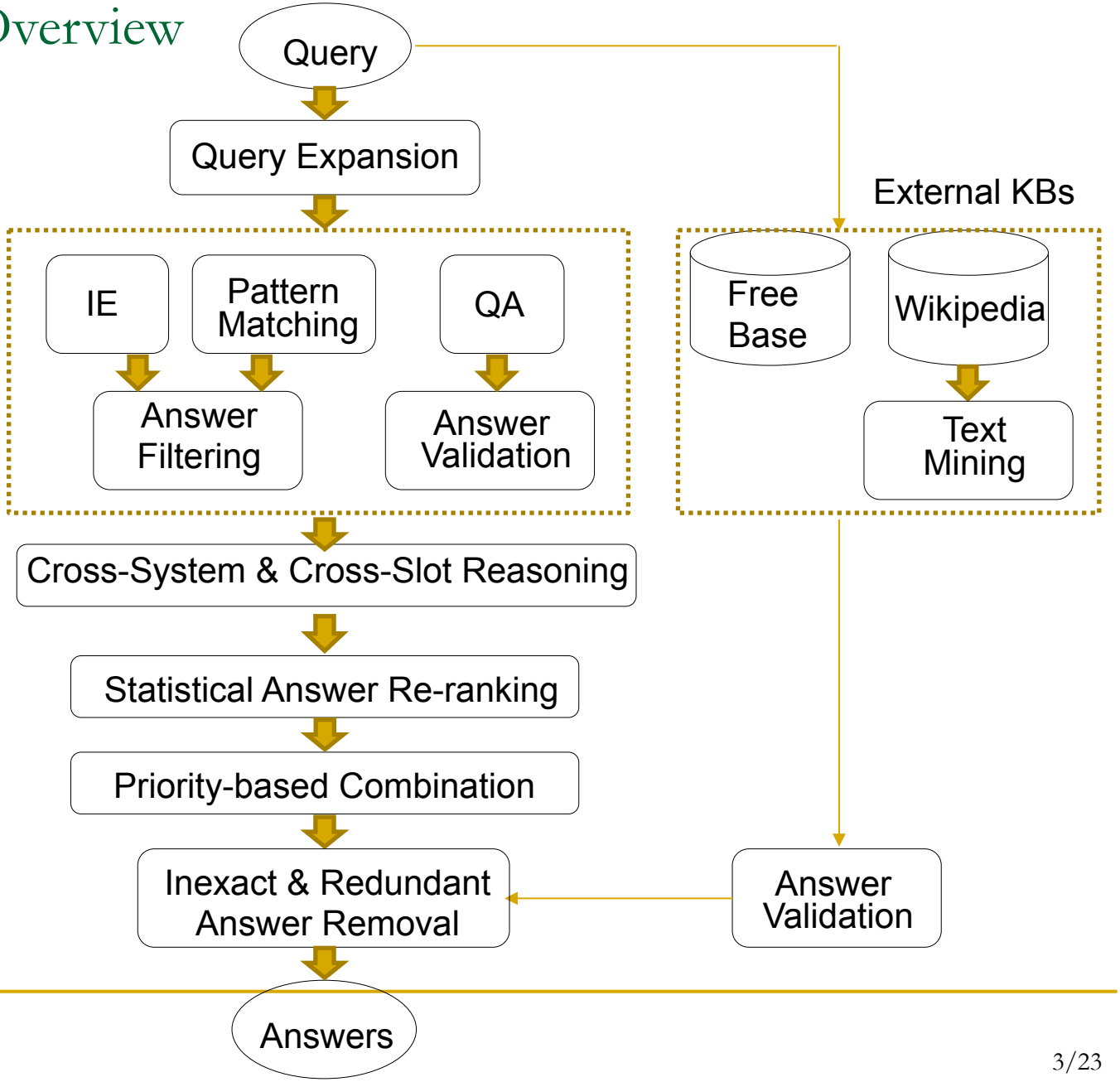
Computer Science Department
Queens College and the Graduate Center
City University of New York

hengji@cs.qc.cuny.edu
November 15, 2010

Outline

- Participated in regular EL, regular SF and surprise SF
- Slot Filling
  - Pipelines
    - Information Extraction
    - Pattern Learning
    - Question-Answering
  - External Knowledge Bases
  - Cross-Slot Reasoning
  - Statistical Re-Ranking
  - Experimental Results
  - Active Learning via Statistical Re-Ranking
- Entity Linking
  - Enhanced by Entity Profiling based on Slot Filling Feedback

System Overview



IE Pipeline

- Apply ACE Cross-document IE (Ji et al., 2009)
- Mapping ACE to KBP, examples:

KBP 2010 slots	ACE2005 relations/ events
per:date_of_birth, per:country_of_birth, per:stateorprovince_of_birth, per:city_of_birth	*event: be-born*
per:countries_of_residence, per:stateorprovinces_of_residence, per:cities_of_residence,per:religion	*relation:citizen-resident-religion-ethnicity*
per:school_attended	*relation:student-alum*
per:member_of	*relation:membership, relation:sports-affiliation*
per:employee_of	*relation:employment*
per:spouse, per:children, per:parents, per:siblings, per:other_family	*relation:family, event: marry, event:divorce*
per:charges	*event:charge-indict, event:convict*

Pattern Learning Pipeline

- Selection of query-answer pairs from Wikipedia Infobox
  - split into two sets
- Pattern extraction
  - For each {*q,a*} pair, generalize patterns by entity tagging and regular expressions e.g. <*q*> died at the age of <*a*>
- Pattern assessment
  - Evaluate and filter based on matching rate
- Pattern matching
  - Combine with coreference resolution
- Answer Filtering based on entity type checking, dictionary checking and dependency parsing constraint filtering

QA Pipeline

- Apply open domain QA system, OpenEphyra (Schlaefer et al., 2007)
- Relevance metric related to PMI and CCP
  - Answer pattern probability:

    P (q, a) = P (q NEAR a): NEAR within the same sentence boundary

$$R(q,a) = \frac{freq(q\ NEAR\ a)}{freq(q) \times freq(a)} \times \#sentences$$

- Limited by occurrence based confidence and recall issues

More Queries and Fewer Answers

- Query Template expansion
  - Generated 68 question templates for organizations and 68 persons
    - Who founded <org>?
    - Who established <org>?
    - <org> was created by who?
- Query Name expansion
  - Wikipedia redirect links
- Heuristic rules for Answer Filtering
  - Format validation
  - Gazetteer based validation
  - Regular expression based filtering
  - Structured data identification and answer filtering

Motivation of Statistical Re-Ranking

- Union and voting are too sensitive to the performance of baseline systems
  - Union guarantees highest recall
    - requires comparable performance
  - Voting
    - assumes more frequent answers are more likely true (FALSE)
  - Priority-based combination
    - voting with weights
    - assumes system performance does not vary by slot (FALSE)

Slot	IE	QA	PL
org:country_of_headquarters	**75.0**	15.8	16.7
org:founded	-	**46.2**	-
per:date_of_birth	100	33.3	**76.9**
per:origin	-	22.6	**40**

Statistical Re-Ranking

- Maximum Entropy (MaxEnt) based supervised re-ranking model to re-rank candidate answers for the same slot
- Features
  - Baseline Confidence
  - Answer Name Type
  - Slot Type X System
  - Number of Tokens X Slot Type
  - Gazetteer constraints
  - Data format
  - Context sentence annotation (dependency parsing, …)
  - …

MLN-based Cross-Slot Reasoning

- Motivation
  - each slot is often dependent on other slots
  - can construct new 'revertible' queries to verify candidate answers
  - *X* is *per:children* of *Y* → *Y* is *per:parents* of *X*;
  - *X* was born on date *Y* → age of *X* is approximately (the current year – *Y*)

- Use Markov Logic Networks (MLN) to encode cross-slot reasoning rules
  - Heuristic inferences are highly dependent on the order of applying rules
  - MLN can
    - adds a weight to each inference rule
    - integrates soft rules and hard rules

External Knowledge Base Construction, Search and Validation

- Used to discover new and validate existing answers
- Freebase - harvests information for many open data sets
  - Information is categorized by criteria; *"American Football"* includes the divisions coach, player, statistics, etc.
  - Relevant criteria were mapped to KBP slots
- Wikipedia text mining as offline KB
  - was used to increase answer confidence
  - If *{q,a}* appeared in the same Wiki, conf increased to 1.

Impact of Statistical Re-Ranking

	Pipelines	Precision	Recall	F-measure
Bottom-up	Supervised IE	0.2416	0.1421	0.1789
	Pattern Matching	0.2186	0.3769	0.2767
Top-down	QA	0.2668	0.1730	0.2099
Priority based Combination		0.3048	0.2658	0.2840
Re-Ranking based Combination		0.2797	0.4433	**0.3430**

- 5-fold cross-validation on training set
- Mitigate the impact of errors produced by scoring based on co-occurrence (slot type x sys feature)
- e.g. the query "*Moro National Liberation Front*" and answer "*1976*"did not have a high co-occurrence, but was bumped up by the re-ranker based on the slot type feature *org:founded*

Impact of Cross-Slot Reasoning

Operations	Total	Correct(%)	Incorrect(%)
Removal	277	88%	12%
Adding	16	100%	0%

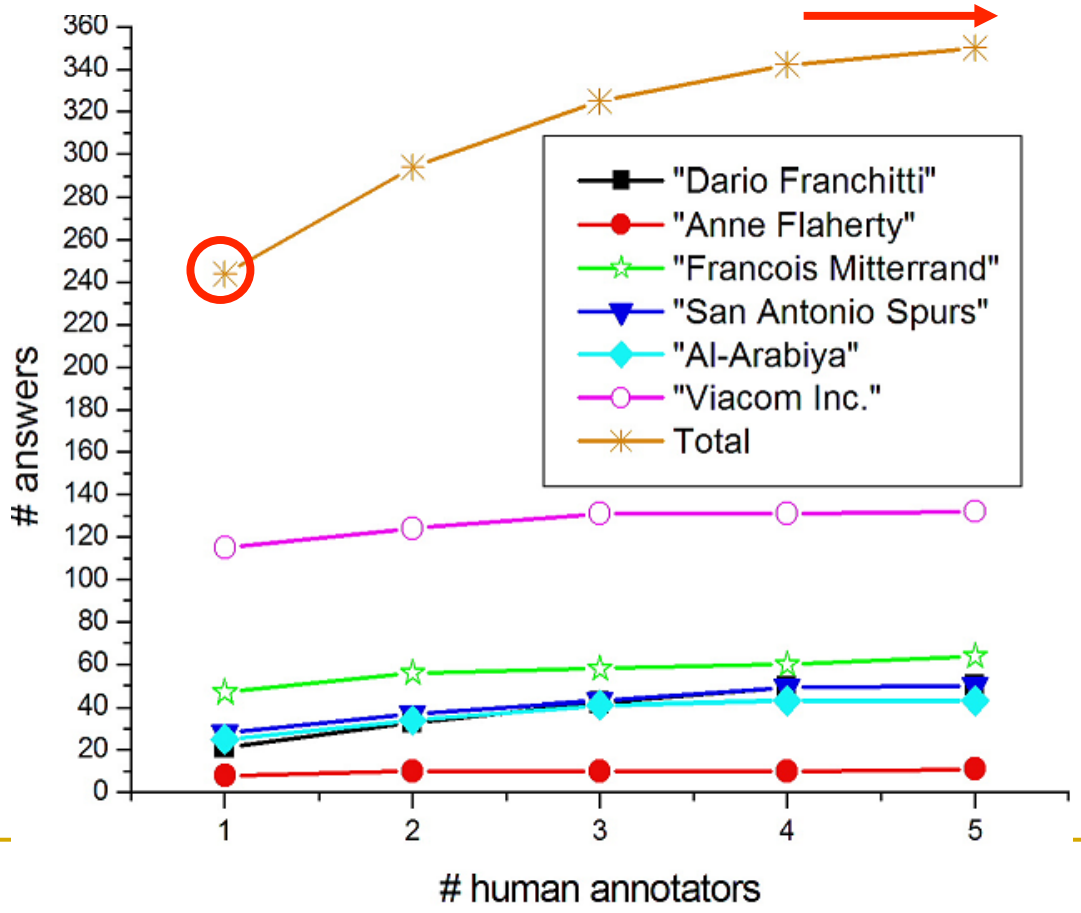
Brian McFadden | per:title | singers | "She had two daughter with one of the MK'd Westlife singers, Brian McFadden, calling them Molly Marie and Lilly Sue"

Impact of Using External Knowledge/ Wikipedia Text Mining

Using Freebase/Wiki Text Mining	Precision	Recall	F-measure
NO	27.99%	26.02%	26.97%
YES	28.74%	27.85%	28.29%

Slot Filling Annotation Bottleneck

- The overlap rates between two participant annotators in community are generally lower than 30%
- Keep adding more human annotators help? No



Can Amazon Mechanical Turk Help?

amazonmechanical turk beta
Artificial Artificial Intelligence

- Given a q, a and supporting context sentence, Turk should judge if the answer is
  - Y: correct; N: incorrect; U: unsure
- Result Distribution for 1690 instances

Useful Annotations (41.8%)		Useless Annotations (58.2%)	
Cases	**Number**	**Cases**	**Number**
Y Y Y Y Y	230	Y Y Y N N	164
N N N N N	16	Y Y Y N U	165
Y Y Y Y U	151	N N N Y Y	158
N N N N U	24	Y Y N N U	171
Y Y Y Y N	227	Y Y N U U	77
N N N N Y	46	Y Y U U U	17
Y Y Y U U	13	N N N Y U	72
N N N U U	59	N N Y U U	57
		Y N U U U	22
		Y U U U U	8
		N N U U U	11
		N U U U U	1
		U U U U U	1

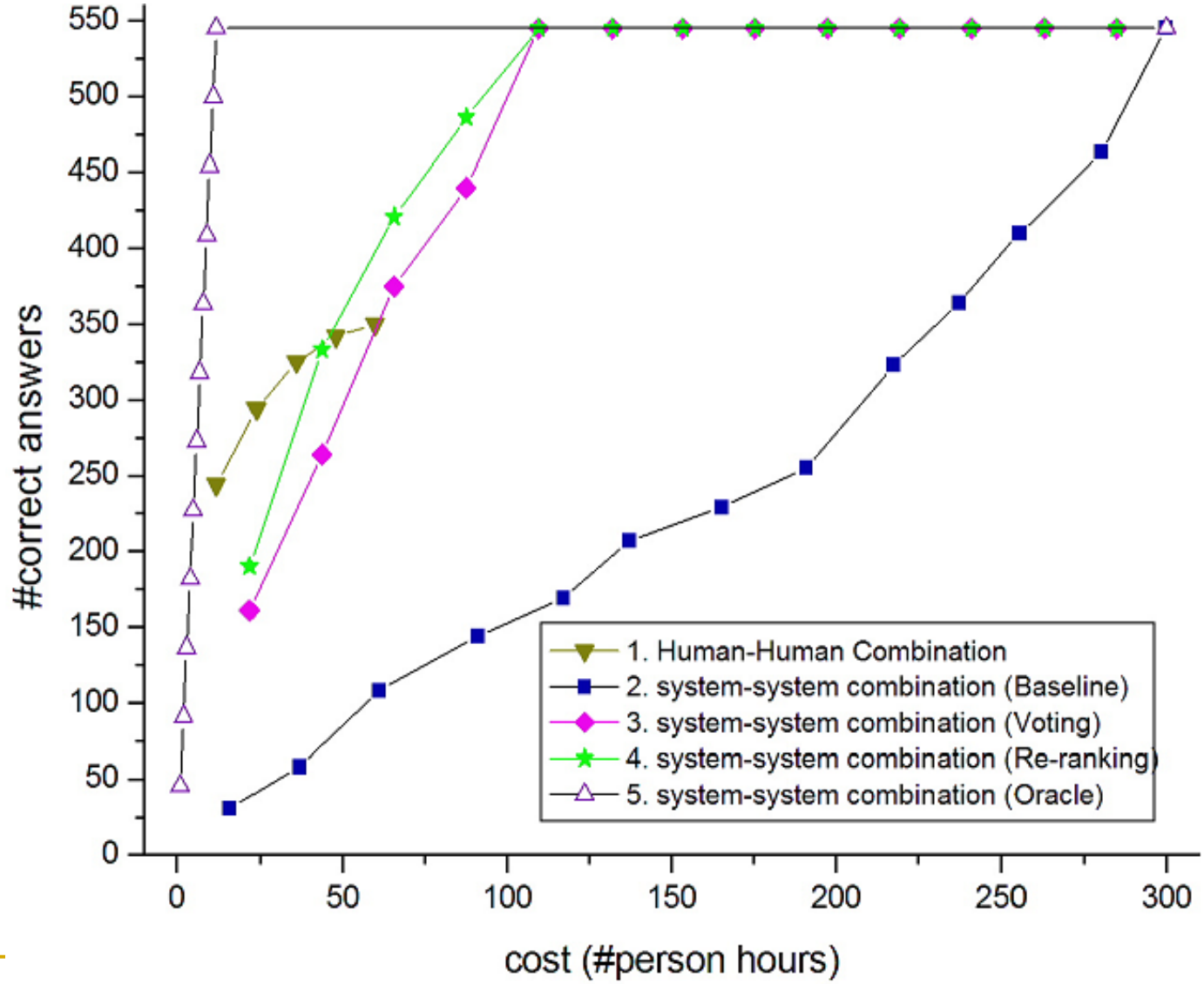
Why is Annotation so hard for Non-Experts?

- Even for all-agreed cases, some annotations are incorrect…

Query	Slot	Answer	Context
Citibank	org: top_members /employees	**Tim Sullivan**	He and **Tim Sullivan**, **Citibank**'s Boston area manager, said they still to plan seek advice from activists going forward.
International Monetary Fund	org: subsidiaries	**World Bank**	President George W. Bush said Saturday that a summit of world leaders agreed to make reforms to the **World Bank** and **International Monetary Fund**.

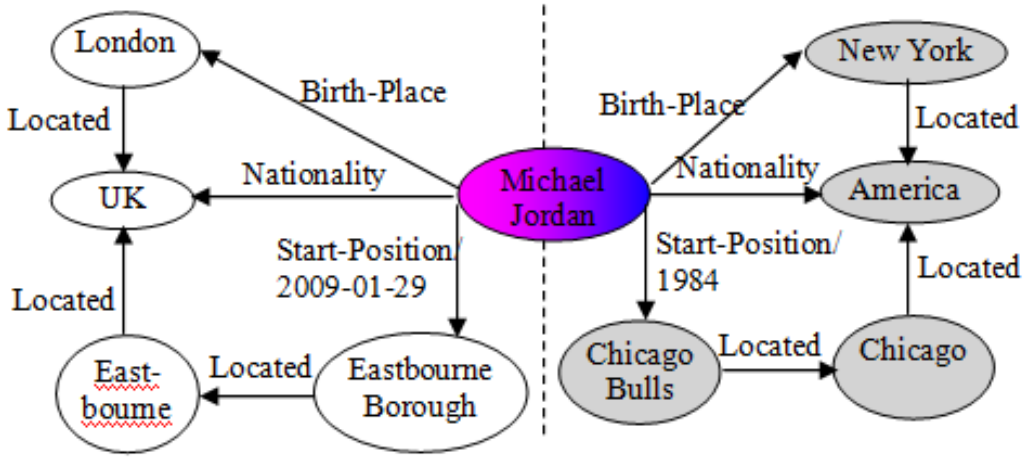
- Require quality control
- Training difficulties

Statistical Re-Ranking based Active Learning

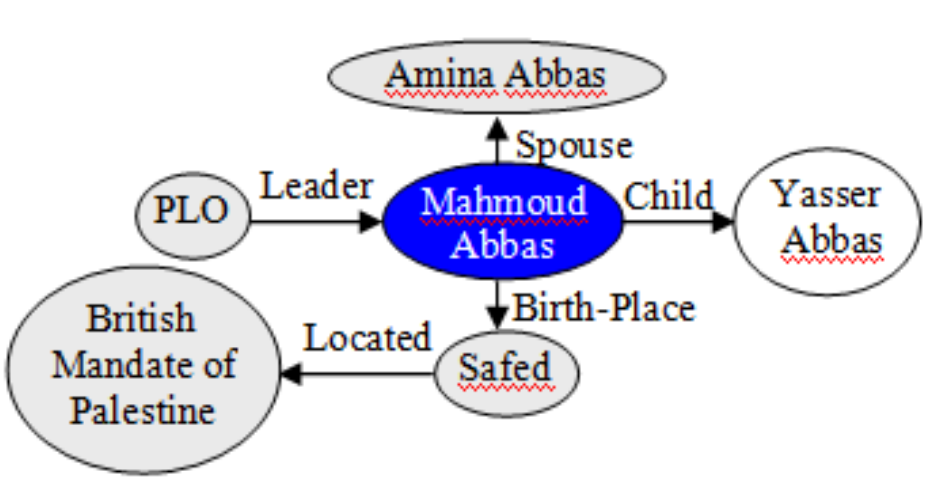


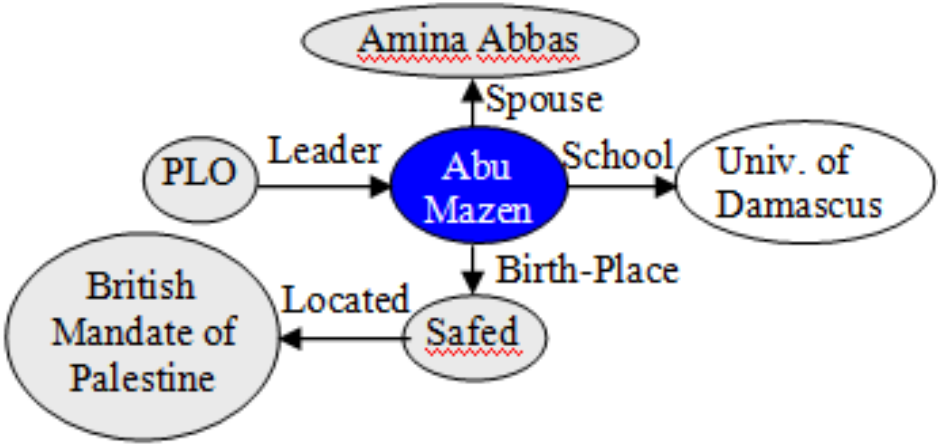
Using Slot Filling Feedback for Entity Linking

Disambiguation



Name Variant Clustering





Using Slot Filling Feedback for Entity Linking

- Two approaches
  - Baseline: Vector space model (tf-idf)
  - Adding entity profile (birth-date, title etc.) from slot filling as additional features
- Results

System	Person	Organization	Overall
Without SF feedback	84.6%	63.1%	59.9%
With SF feedback	92.8%	65.7%	69.3%

Conclusions

- Big gains from statistical re-ranking combining 3 pipelines
  - Information Extraction
  - Pattern Learning
  - Question-Answering
- Further gains from MLN cross-slot reasoning
- Automatic profiles from SF dramatically improve EL
- Human-system combination provides efficient answer-key generation
  - Faster, better, cheaper!