Customizing an Information Extraction System to a New Domain

Mihai Surdeanu, David McClosky, Mason R. Smith, Andrey Gusev, and Christopher D. Manning

The Scenario

- The task
  - Build an IE system (entity mention and *relation mention* detection – EMD, RMD) for a domain you have not seen before
  - You have to deliver under a tight deadline

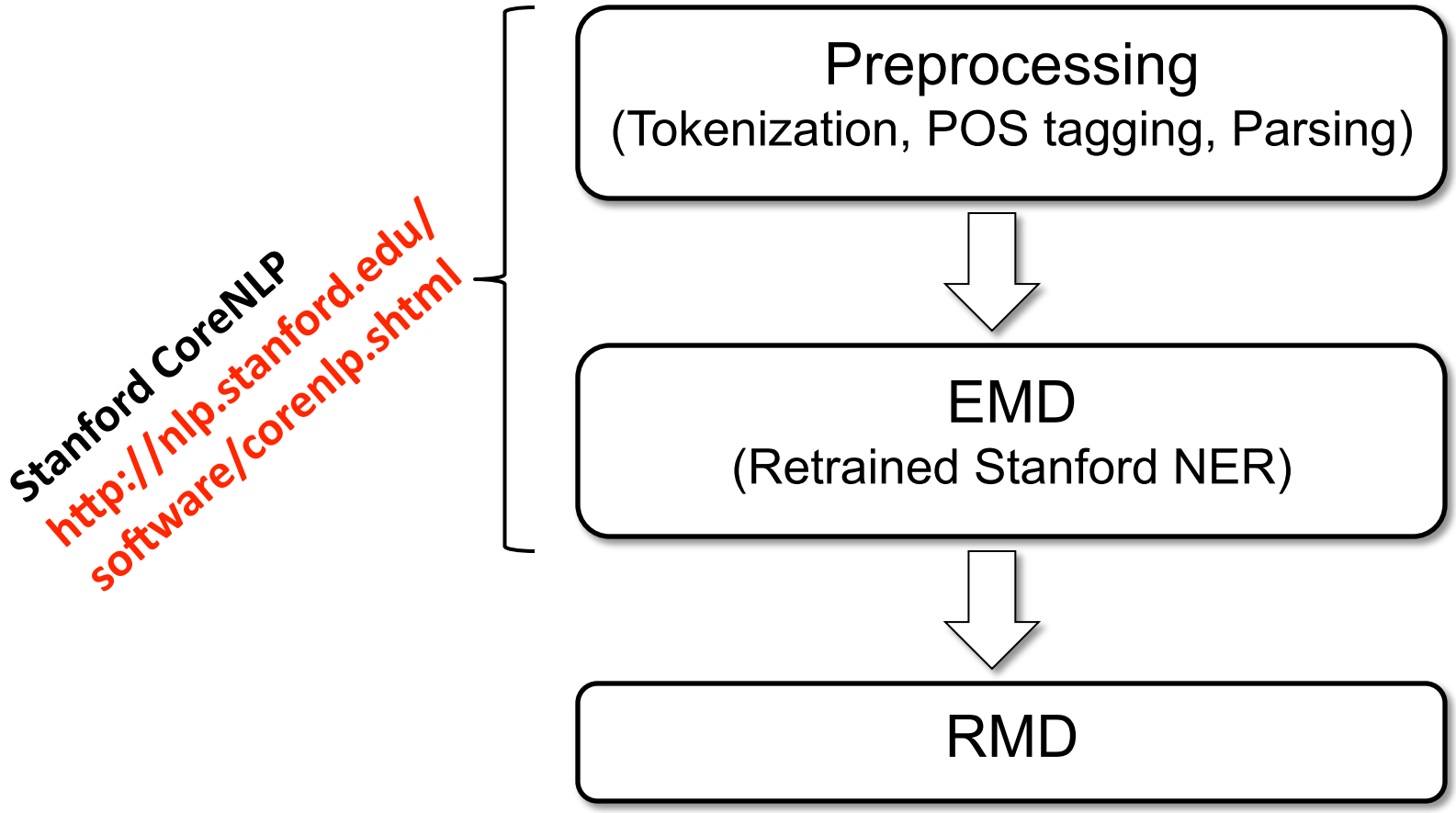
- The input
  - Training data for the new domain
  - An existing EMD + RMD system developed for a ”classic” domain – ACE

A Story of Domain Customization

- Will focus on several simple ideas that are usually overlooked during domain customization

- What we do not do (but is complementary)
  - Feature design for the new domain
  - Domain adaptation models

Our IE System



Relation Mention Classifier

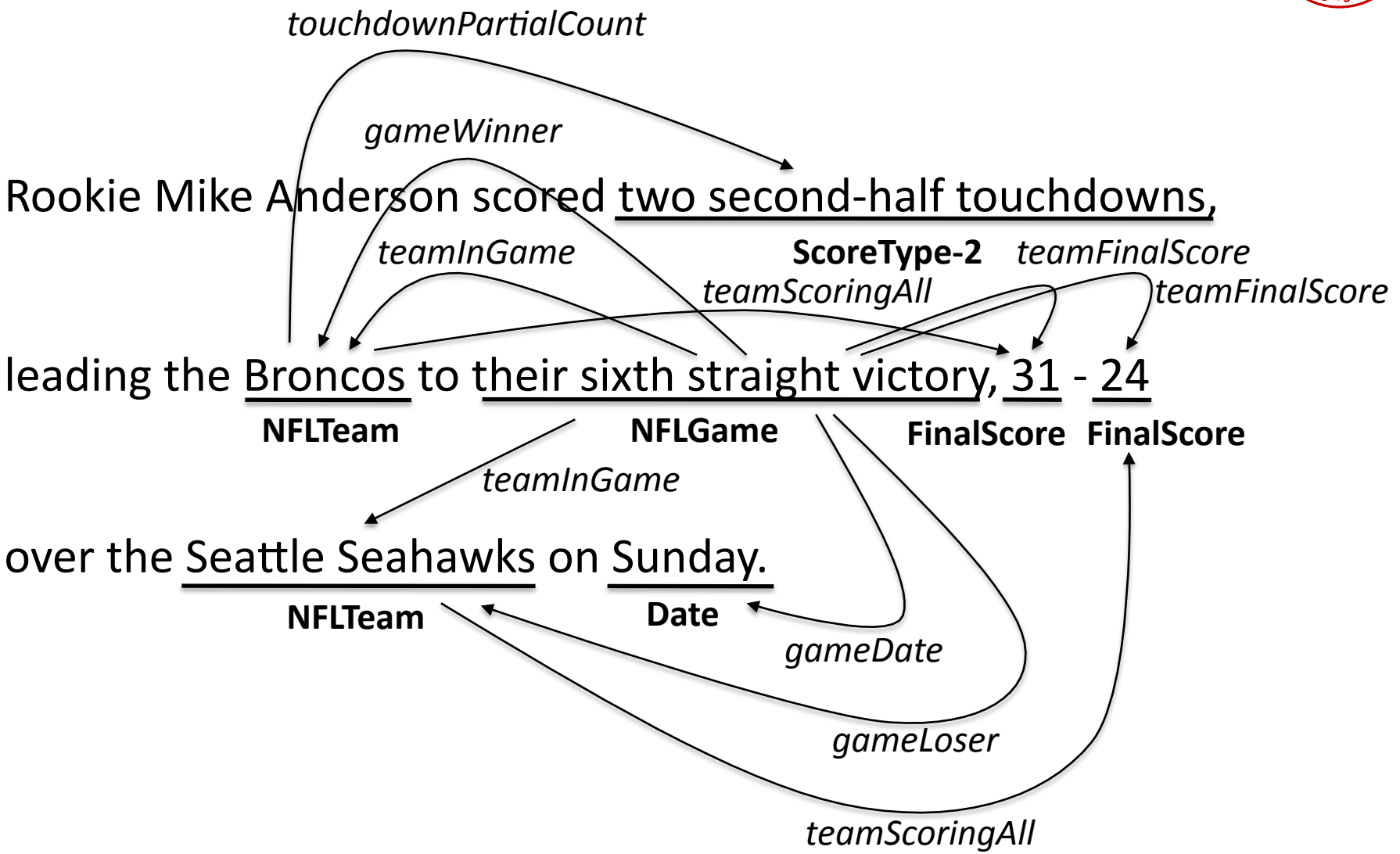
- Multiclass classifier
- Extracts binary relations between entities in the same sentence
- Logistic regression with L2 regularization
- Single label prediction: one relation mention between two entity mentions

RMD Features

Argument Features	Head words
	Entity labels
Syntactic Features	Dependency path between argument heads
	Lemmas of words in path
	Syntactic path in constituent tree between argument heads
Surface Distance Features	POS tags between arguments
	Entity mentions in between

State of the art performance on ACE 2007 data

The New Domain: NFL Games



The New Domain: NFL Games

- Different from ACE…
  - Some entities may be particularly hard to detect, e.g., NFLGame
  - There is a significant amount of domain knowledge, e.g., NFLTeam
  - Some relations may require inference, e.g., gameWinner, gameLoser
  - Some relations are overlapping, e.g., gameWinner and teamInGame

The New Domain: NFL Games

Documents	Words	Entity Mentions	Relation Mentions
110	70,119	2,188	1,629

Results

	Baseline
Entity Mentions	
Date	77.5
FinalScore	88.6
NFLGame	55.5
NFLTeam	78.7
ScoreType-1	65.5
ScoreType-2	66.7
Overall EMD	**73.7**
Relation Mentions	
gameDate	42.4
gameLoser	26.2
gameWinner	8.7
teamFinalScore	57.1
teamInGame	27.1
teamScoringAll	73.1
Overall RMD	**49.7**

NFLGame is hard

NFLTeam should be better

Hard because they are often implicit

Affected by poor EMD performance

1. Adding Gazetteers

- Constructed a domain gazetteer
  - 32 team names
    - Allow partial matches, e.g., “Cowboys” for “Dallas Cowboys”
  - 50 game descriptors bootstrapped from Dekang Lin’s thesaurus
    - 3 seeds: “victory”, “loss”, “game”
    - 47 new descriptors: “triumph”, “defeat”, etc.

Results

		Baseline	+ gazetteer
Entity Mentions	Date	77.5	
	FinalScore	88.6	+ 0.6
	NFLGame	55.5	- 1.5
	NFLTeam	78.7	+ 0.8
	ScoreType-1	65.5	- 0.1
	ScoreType-2	66.7	
	Overall EMD	**73.7**	**+ 0.3**
Relation Mentions	gameDate	42.4	+ 0.3
	gameLoser	26.2	+ 0.9
	gameWinner	8.7	
	teamFinalScore	57.1	
	teamInGame	27.1	+ 0.2
	teamScoringAll	73.1	+ 1.2
	Overall RMD	**49.7**	**+ 0.5**

2. Model Combination for EMD

- Previous results somewhat disappointing
  - The EMD model favors the LOCATION interpretation for team names represented as city name
  - But we need high recall for RMD…
- Rule-based model for NFLTeam:
  - Token sequence that begins, ends, or equals a gazetteer entry for NFLTeam ➔ NFLTeam
    - For “Green Bay Packers”: “Green Bay”, “Packers”, but not “Bay”
- Combine its output with that of the statistical model

Results

		Baseline	+ gazetteer	+ combo
Entity Mentions	Date	77.5		
	FinalScore	88.6	+ 0.6	
	NFLGame	55.5	- 1.5	
	NFLTeam	78.7	+ 0.8	+ 2.8
	ScoreType-1	65.5	- 0.1	
	ScoreType-2	66.7		
	Overall EMD	**73.7**	**+ 0.3**	**+ 1.5**
Relation Mentions	gameDate	42.4	+ 0.3	- 0.8
	gameLoser	26.2	+ 0.9	+ 1.0
	gameWinner	8.7		+ 1.2
	teamFinalScore	57.1		
	teamInGame	27.1	+ 0.2	+ 5.0
	teamScoringAll	73.1	+ 1.2	+ 2.6
	Overall RMD	**49.7**	**+ 0.5**	**+ 3.0**

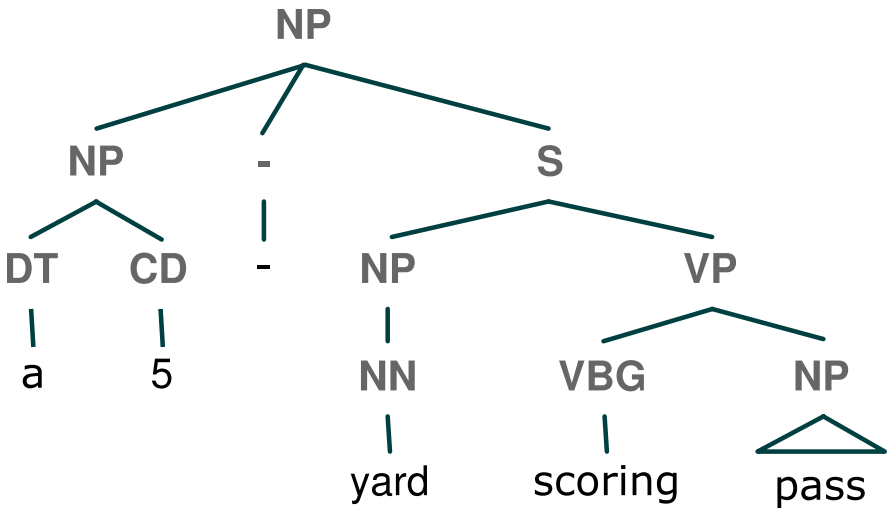
3. Better Identification of Syntactic Heads

- Why do we care?

Argument Features	**Head** words
	Entity labels
Syntactic Features	Dependency path between argument **heads**
	Lemmas of words in path between **heads**
	Syntactic path in constituent tree between argument **heads**
Surface Distance Features	POS tags between arguments
	Entity mentions in between

3. Better Identification of Syntactic Heads

- "Classic" head finding heuristic:
  - Try to find a constituent with the same span in the constituent tree and extract its head
    - But more than 25% of mentions cannot be matched to a constituent…
  - If none found, parse the text of the standalone mention
    - But parsing short out-of-domain text is hard



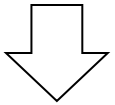
Head word: "5"

3. Better Identification of Syntactic Heads

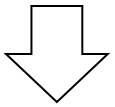
- Proposed heuristic:
  - Try to find a constituent with the same span in the constituent tree and extract its head
  - If none found, parse the mention:
    - Append “It was” to beginning (so it looks like a sentence)
    - Remove dashes from text (not common in original Treebank)
    - Force the parser to generate a constituent with the same span as the mention

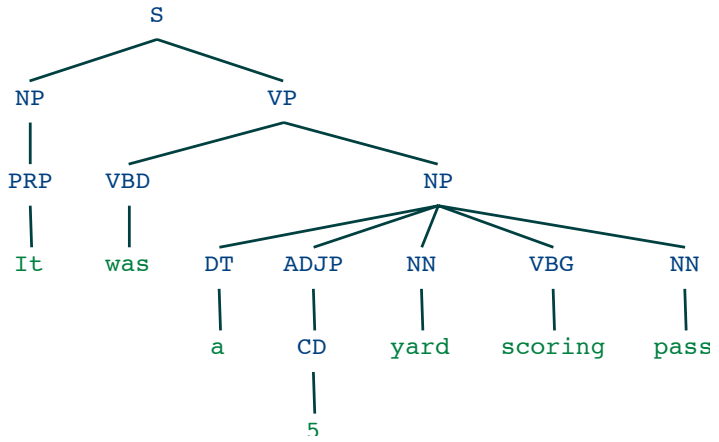
3. Better Identification of Syntactic Heads

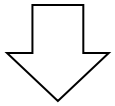
a 5 – yard scoring pass



It was a 5 yard scoring pass





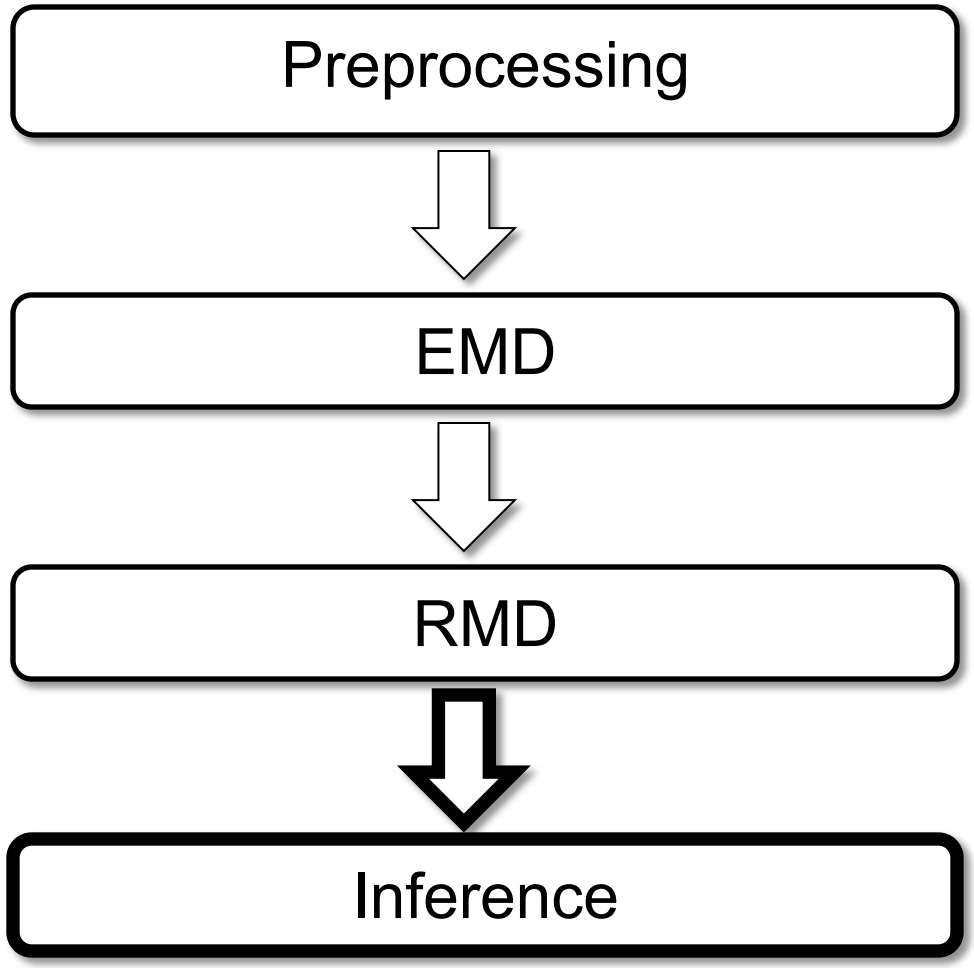


Head word: “pass”

Results

		Baseline	+ gazetteer	+ combo	+ head
Entity Mentions	Date	77.5			- 5.0
	FinalScore	88.6	+ 0.6		+ 0.6
	NFLGame	55.5	- 1.5		+ 1.4
	NFLTeam	78.7	+ 0.8	+ 2.8	+ 1.1
	ScoreType-1	65.5	- 0.1		+ 0.3
	ScoreType-2	66.7			+ 2.7
	Overall EMD	**73.7**	**+ 0.3**	**+ 1.5**	**+ 0.6**
Relation Mentions	gameDate	42.4	+ 0.3	- 0.8	
	gameLoser	26.2	+ 0.9	+ 1.0	- 1.8
	gameWinner	8.7		+ 1.2	+ 5.3
	teamFinalScore	57.1			+ 6.5
	teamInGame	27.1	+ 0.2	+ 5.0	+ 7.2
	teamScoringAll	73.1	+ 1.2	+ 2.6	+ 2.3
	Overall RMD	**49.7**	**+ 0.5**	**+ 3.0**	**+ 4.7**

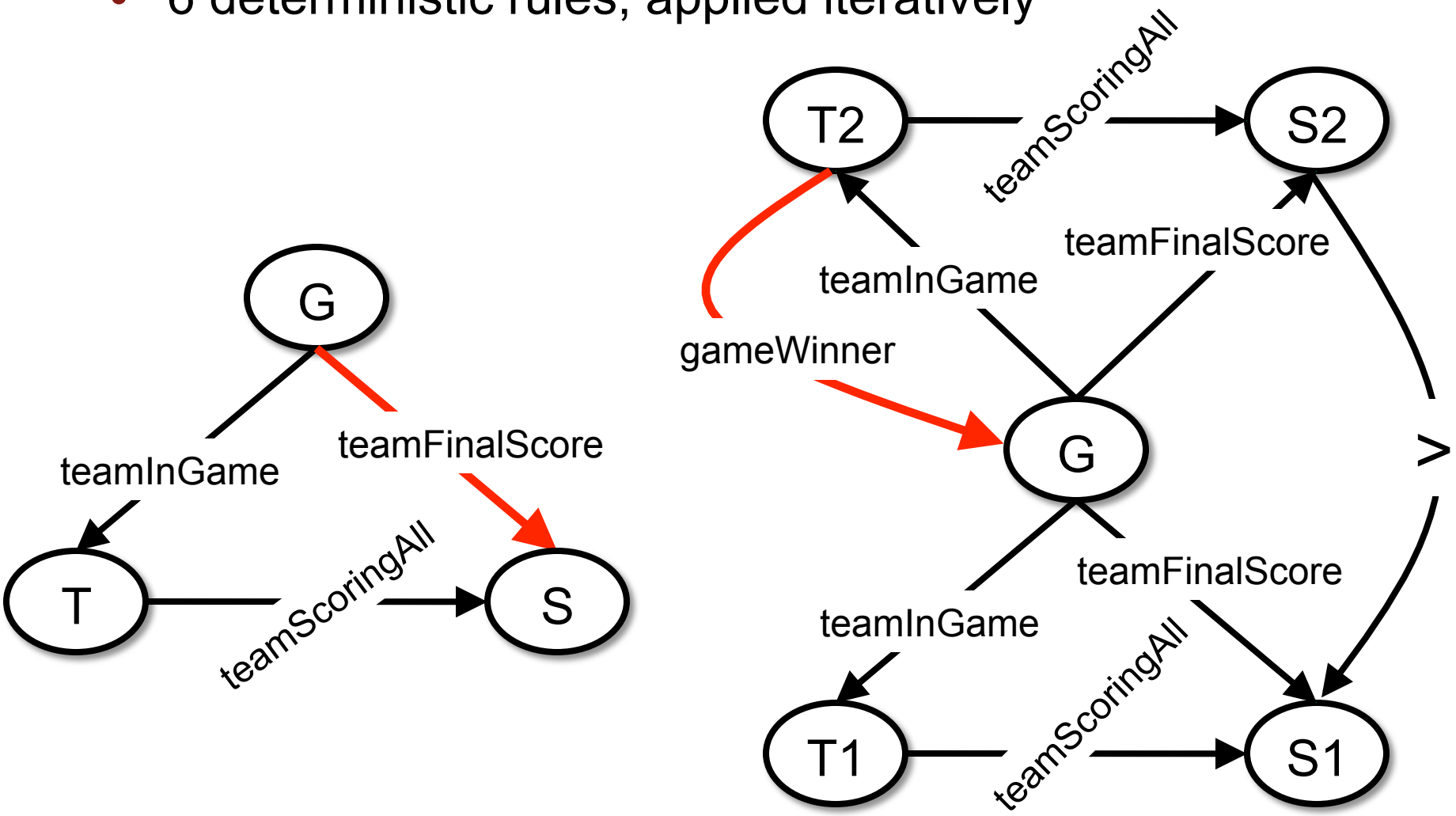
4. Deterministic Inference



Generates new relations based on data already extracted

4. Deterministic Inference

- 6 deterministic rules, applied iteratively



Results

		Baseline	+ gazetteer	+ combo	+ head	+ inference
Entity Mentions	Date	77.5			- 5.0	
	FinalScore	88.6	+ 0.6		+ 0.6	
	NFLGame	55.5	- 1.5		+ 1.4	
	NFLTeam	78.7	+ 0.8	+ 2.8	+ 1.1	
	ScoreType-1	65.5	- 0.1		+ 0.3	
	ScoreType-2	66.7			+ 2.7	
	Overall EMD	**73.7**	**+ 0.3**	**+ 1.5**	**+ 0.6**	
Relation Mentions	gameDate	42.4	+ 0.3	- 0.8		
	gameLoser	26.2	+ 0.9	+ 1.0	- 1.8	+ 8.0
	gameWinner	8.7		+ 1.2	+ 5.3	+ 17.3
	teamFinalScore	57.1			+ 6.5	
	teamInGame	27.1	+ 0.2	+ 5.0	+ 7.2	+ 6.9
	teamScoringAll	73.1	+ 1.2	+ 2.6	+ 2.3	
	Overall RMD	**49.7**	**+ 0.5**	**+ 3.0**	**+ 4.7**	**+ 1.6**

Stress Testing the Inference

	Without inference	With inference
Skip gameWinner, gameLoser	57.7	58.8
Skip teamInGame	55.7	58.5
Skip teamInGame, teamFinalScore	49.6	56.9
Skip nothing	57.9	**59.5**

Summary

	Entity Mentions	Relation Mentions
Baseline	73.7	49.7
+ gazetteer features	74.0	50.2
+ model combination for NFLTeam	75.5	53.2
+ improved head identification	**76.1**	57.9
+ inference	**76.1**	**59.5**

20% relative improvement for RMD

Official Evaluation

- "For each NFL game, identify the winning and losing teams and each team's final score in the game."
- "For each team losing to the Green Bay Packers, tell us the losing team and the number of points they scored."

Needs event coreference

- 50 queries
- 46.7 F1 (53.7 precision and 41.2 recall)

70% of human performance

Conclusions

- Discussed several simple ideas for improved domain customization of EMD + RMD systems
- Observations:
  - The accurate identification of syntactic heads of entity mentions is important
  - Inference can work even for pipeline, non-global systems
  - Combining statistical models with rule-based models helps
- 20% relative improvement in RMD F1 score over a state-of-the-art system

THANK YOU!