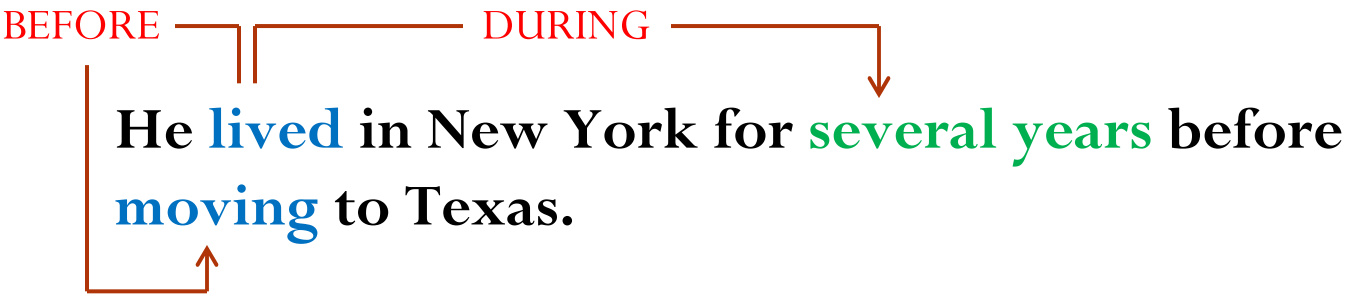
Classifying Temporal Relations with Rich Linguistic Knowledge

Jennifer D'Souza and Vincent Ng
Human Language Technology Research Institute
The University of Texas at Dallas

Task Definition

- Given two **entities** (i.e. events or time expressions) in a text document classify them into one of a set of predefined temporal relations.



BEFORE

DURING

He lived in New York for several years before moving to Texas.

Goal

- Advance the state of the art in temporal relation classification by working on a more complex version of the task.
    - Attempt 14-class classification as opposed to the typical 3 or 6 class classification task (Mani et al., (2006), Chambers et al. (2007), Verhagen et al., (2007), Verhagen et al., (2010), Mirroshandel and Ghaseem-Sani (2011))

Our Approach

- Knowledge-rich
  - large scale expansion of linguistic features
    - **semantic** and **discourse** features
  - other approaches have relied on primarily morpho-syntactic features
- Hybrid
  - propose a system architecture in which we combine **learning-based** approach and **rule-based** approach
    - other approaches are either learning-based or rule-based
  - **Hypothesis**: rule-based method could better handle
    - skewed class distribution
    - leverage human insights to combine linguistic features

Plan for the talk

- Dataset
- Baseline Temporal Relation Classifier
- Our Knowledge-Rich, Hybrid Approach
  - Novel Features
  - Combining Rules and Machine Learning
- Evaluation

Plan for the talk

- Dataset
- Baseline Temporal Relation Classifier
- Our Knowledge-Rich, Hybrid Approach
  - Novel Features
  - Combining Rules and Machine Learning
- Evaluation

Dataset

- TimeBank
  - 183 newswire articles annotated with 14 temporal relations

14 types of event-event, event-time temporal relations

Simultaneous (13.3%)	**Identity (14.1%)**
Before (13.9%)	**After (15%)**
Includes (15.3%)	**Is_Included (15.3%)**
IBefore (0.8%)	**IAfter (0.6%)**
During (2.1%)	**During_Inv (2.5%)**
Begins (1.3%)	**Begun_By (1.2%)**
Ends (1.3%)	**Ended_By (3.4%)**

Dataset

- TimeBank
  - 183 newswire articles annotated with 14 temporal relations

14 types of event-event, event-time temporal relations

Simultaneous (13.3%)	**Identity (14.1%)**
Before (13.9%)	**After (15%)**
Includes (15.3%)	**Is_Included (15.3%)**
IBefore (0.8%)	**IAfter (0.6%)**
During (2.1%)	**During_Inv (2.5%)**
Begins (1.3%)	**Begun_By (1.2%)**
Ends (1.3%)	**Ended_By (3.4%)**

Dataset

- TimeBank
  - 183 newswire articles annotated with 14 temporal relations

14 types of event-event, event-time temporal relations

Simultaneous (13.3%)	**Identity (14.1%)**
Before (13.9%)	**After (15%)**
Includes (15.3%)	**Is_Included (15.3%)**
IBefore (0.8%)	**IAfter (0.6%)**
During (2.1%)	**During_Inv (2.5%)**
Begins (1.3%)	**Begun_By (1.2%)**
Ends (1.3%)	**Ended_By (3.4%)**

Plan for the talk

- Dataset
- Baseline Temporal Relation Classifier
- Our Knowledge-Rich, Hybrid Approach
  - Novel Features
  - Combining Rules and Machine Learning
- Evaluation

Learning-based Baseline Temporal Relation Classifier

- Training Instance Creation
  - Each instance corresponds to two entities (entity1,entity2)
    - Class value is one of the 14 temporal relation types

- Conditions to form a training instance:
  - entity1 precedes entity2 in the associated text
  - (entity1,entity2) belongs to one of the 14 temporal relation types

Learning-based Baseline Temporal Relation Classifier

- 62 features
  1. Lexical (5) – based on the entity string
  2. Grammatical (33) – based on grammatical syntax including POS and phrase information
  3. Entity attributes (13) – encode tense, aspect, modality, polarity, and type of event, or type of time expression
  4. Semantic (7) – based on related temporal arguments, WordNet synsets, and VerbOcean relations
  5. Distance (1)
  6. Document Creation Time related (3)

Mani et al. (2006), Chambers et al. (2007), Min et al. (2007), Puscasu (2007), Ha et al. (2010), Llorens et al. (2010), Mirroshandel and Ghassem-Sani (2011)

Learning-based Baseline Temporal Relation Classifier

- $SVM^{multiclass}$ (Tsochantaridis et al., 2004)

Plan for the Talk

- Dataset
- Baseline Temporal Relation Classifier
- Our Knowledge-Rich, Hybrid Approach
  - Novel Features
  - Combining Rules and Machine Learning
- Evaluation

Novel Features

- Five types:
  1. Pairwise Features
  2. Dependency Relation Features
  3. Webster and WordNet Relation Features
  4. Predicate-Argument Relation Features
  5. Discourse Relation Features

Novel Features

- Five types:
  1. Pairwise Features
  2. Dependency Relation Features
  3. Webster and WordNet Relation Features
  4. Predicate-Argument Relation Features
  5. Discourse Relation Features

Pairwise Features

- **Hypothesis:**
  - pairwise features, which are computed based on both entities, could better capture the relation between them. This is missing in some of our features in the baseline.

Pairwise Features

1. Class and tense of entity1 with class and tense of entity2
   - E.g.: … says it will offer …
   - says BEFORE offer
   - Feature value: REPORTING$_1$-PRESENT$_1$-OCCURRENCE$_2$-FUTURE$_2$
2. Tense and aspect of entity1 with tense and aspect of entity2
   - E.g.: The embargo is meant to cripple Iraq by cutting off its exports…
   - cripple AFTER cutting
   - Feature value: INFINITIVE$_1$-NONE$_1$-PRESENT$_2$-PROGRESSIVE$_2$

Pairwise Features

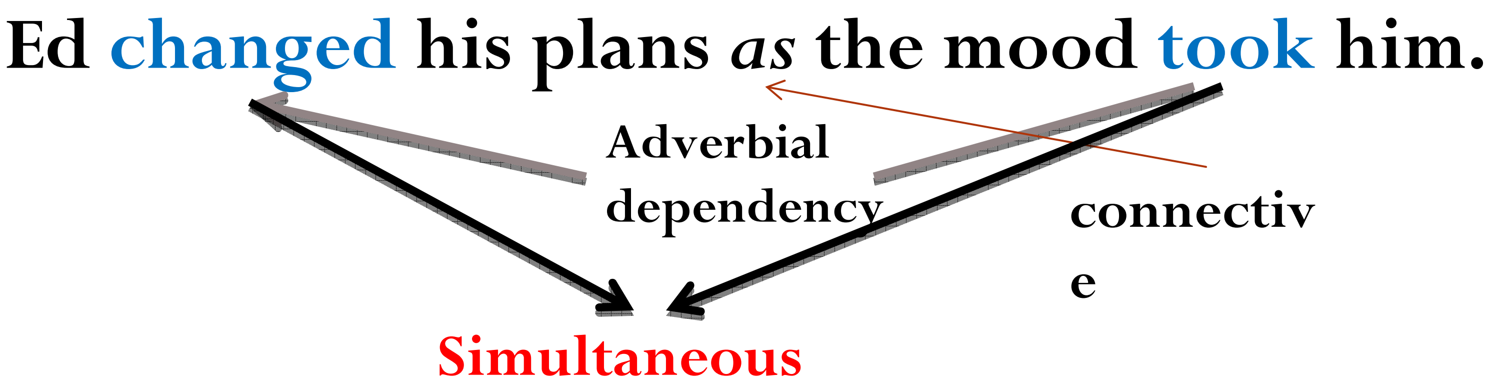
- Some More Pairwise Features

3. Entity head word pairs
4. Prepositional lexeme pairs
5. Preposition trace feature
6. Verb POS trace feature

Novel Features

- Five types:
  1. Pairwise Features
  2. Dependency Relation Features
  3. Webster and WordNet Relation Features
  4. Predicate-Argument Relation Features
  5. Discourse Relation Features

Why are Dependency Relations useful for Temporal Relation Classification?



Ed **<span>changed</span> his plans** ***as*** **the mood took him.**

- adverbial clause dependency relations along with the subordinating conjunction generally inform **Simultaneous**, **Before** or **Aft**er temporal relations

Hypothesis: other types of dependency relations would also be useful for temporal relation classification.

Dependency Relation Features

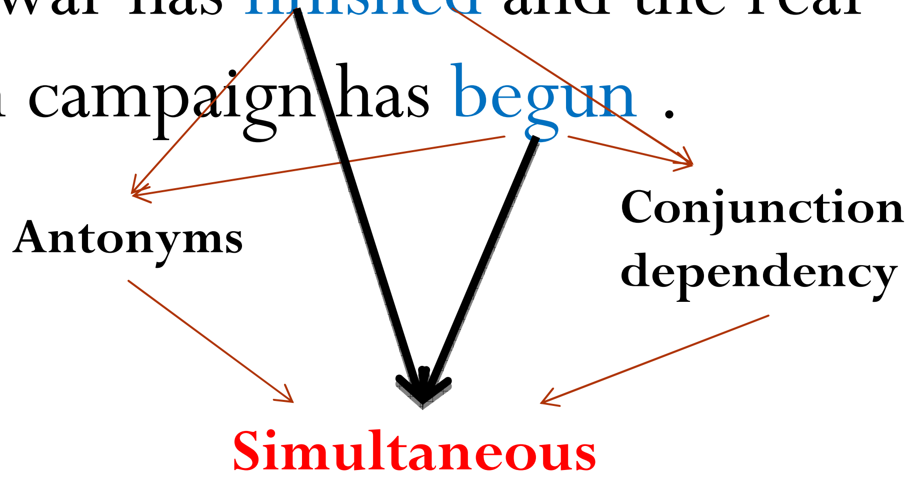
- For each of the 25 dependency relation types produced by the Stanford parser:
  - Is entity1/entity2 the governor in the relation?
  - Is entity1/entity2 the dependent in the relation?

Novel Features

- Five types:
  1. Pairwise Features
  2. Dependency Relation Features
  3. Webster and WordNet Relation Features
  4. Predicate-Argument Relation Features
  5. Discourse Relation Features

Why are Semantic Relations useful for Temporal Relation Classification?

The phony war has finished and the real referendum campaign has begun .



Antonyms

Conjunction dependency

Simultaneous

Hypothesis: other types of semantic relations would also be useful for temporal relation classification.

Webster Relation Features

- 4 types of Webster semantic relations:
  - synonym, related-word, near-antonym, and antonym
- 8 features:
  - for each type of semantic relation t:
    - is (event1, event2)$\in$ t?
    - is (event2, event1)$\in$ t?

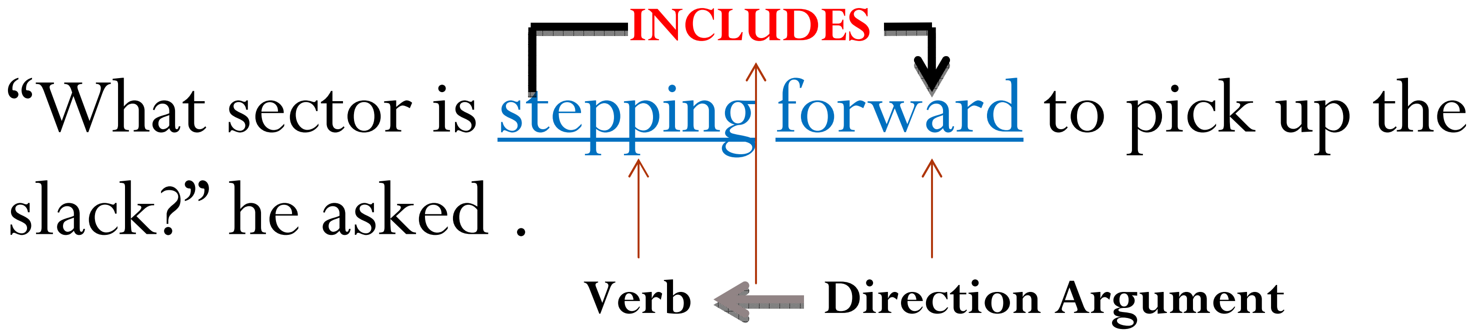
WordNet Relation Features

- 4 types of WordNet semantic relations:
  - hypernym, hyponym, troponym, and similar
- 8 features:
  - for each type of semantic relation:
    - is (event1, event2)ϵ t?
    - is (event2, event1)ϵ t?

Novel Features

- Five types:
  1. Pairwise Features
  2. Dependency Relation Features
  3. Webster and WordNet Relation Features
  4. Predicate-Argument Relation Features
  5. Discourse Relation Features

Why are Predicate-Argument Relations useful for Temporal Relation Classification?



“What sector is stepping forward to pick up the slack?” he asked .

INCLUDES

Verb

Direction Argument

Hypothesis: other types of predicate-argument relations would also be useful for temporal relation classification.

Predicate-Argument Relation Features

- We consider 4 types of predicate-argument relations (obtained automatically using SENNA)
  - directional, manner, temporal and cause
- 8 features:
  - for each type of predicate-argument relation:
    - does event1 appear in event2's argument?
    - does event2 appear in event1's argument?

Novel Features

- Six types:
  1. Pairwise Features
  2. Dependency Relation Features
  3. Webster and WordNet Relation Features
  4. Predicate-Argument Relation Features
  5. Discourse Relation Features

Discourse Relations

- Discourse relations can potentially be exploited to discover both inter-sentential and intra-sentential temporal relations
  - unlike syntactic dependencies and predicate-argument relations, through which we can only identify intra-sentential temporal relations

Why are Discourse Relations useful for Temporal Relation Classification?

Reports **said** that Saudi Arabia told U.S. oil companies of a 15-20 percent cutback in its oil supply in September. Meanwhile Egypt's Middle East Agency said **Thursday** that Saddam was the target of an assassination attempt.

- Explicit Relation: Synchrony
- Intuitively, a reporting event within a discourse unit is_included in a date contained within a separate synchronous discourse unit.

Hypothesis: other types of discourse relations would also be useful for temporal relation classification.

Discourse Relation Features

- 12 types of discourse relations (extracted automatically using Lin et al.'s (2013) PDTB-style discourse parser):
  - Cause, Conjunction, Synchrony, Contrast, …

- 48 features based on explicit discourse relations:
  - for each type of discourse relation:
    - is entity1 in argument1 and entity2 in argument2 of the discourse relation?
    - is entity2 in argument1 and entity1 in argument2 of the discourse relation?

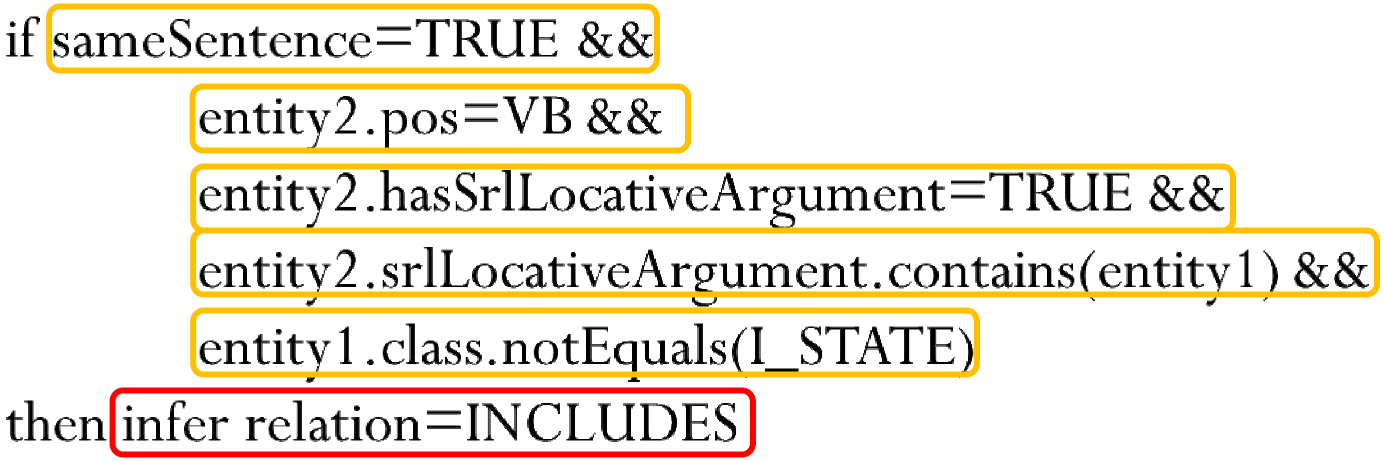
- 48 features based on implicit discourse relations

Plan for the Talk

- Dataset
- Baseline temporal relation classifier
- Our knowledge-rich, hybrid approach
  - Novel features
  - Combining rules and machine learning
- Evaluation

Manual Rule Creation

- The design of rules is partly based on intuition and partly data-driven.
- E.g.



```
if sameSentence=TRUE &&
            entity2.pos=VB &&
            entity2.hasSrlLocativeArgument=TRUE &&
            entity2.srlLocativeArgument.contains(entity1) &&
            entity1.class.notEquals(I_STATE)
then infer relation=INCLUDES
```

Rule Creation and Application

- Rules are manually developed based on development data not used for evaluation.
- Rules are ordered in decreasing order of accuracy measured on development data.
- A new instance is classified using the 1st applicable rule in the ruleset.

Combining Hand-Crafted Rules and Machine Learning

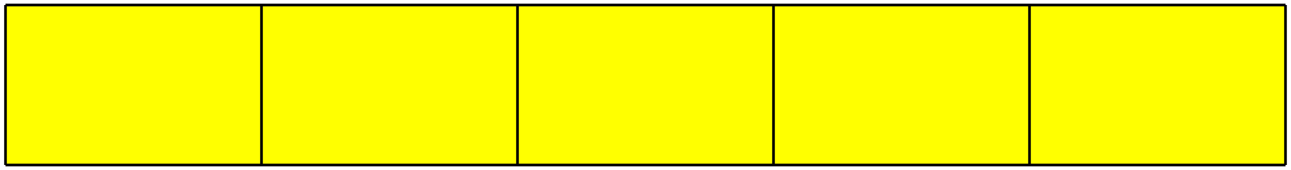
- 3 methods
  - **Method 1**:
    - We employ all of the rules as additional features for training the temporal relation classifier
  - **Method 2**:
    - Given a test instance, we first apply to it the ruleset composed only of rules that are at least 80% accurate. If none of the rules is applicable, we classify it using the classifier employed in method 1.
  - **Method 3**:
    - Same as method 2 except we do not employ the rules as features when training the classifier.

Plan for the Talk

- Dataset
- Baseline temporal relation classifier
- Our knowledge-rich, hybrid approach
  - Novel features
  - Combining rules and machine learning
- Evaluation

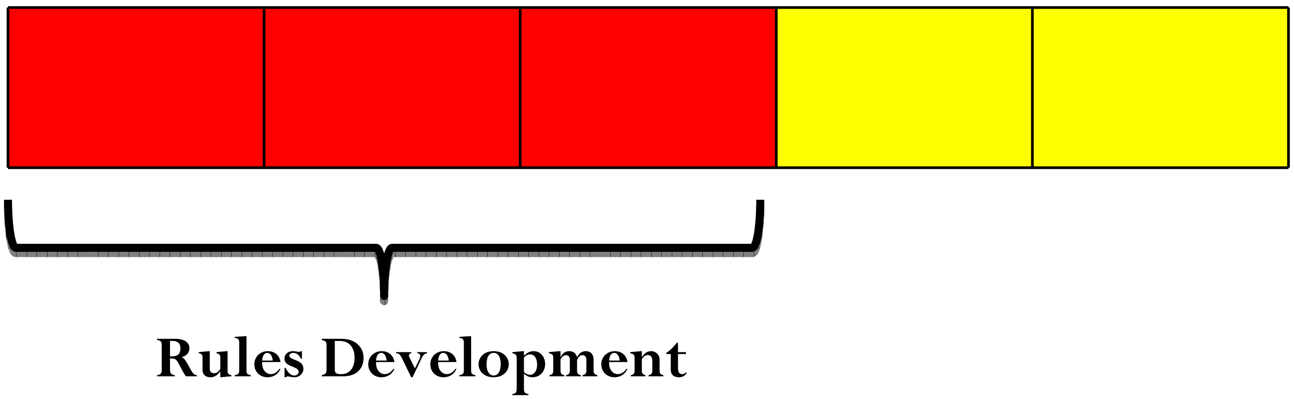
Experimental Setup

- 183 documents in TimeBank



Experimental Setup

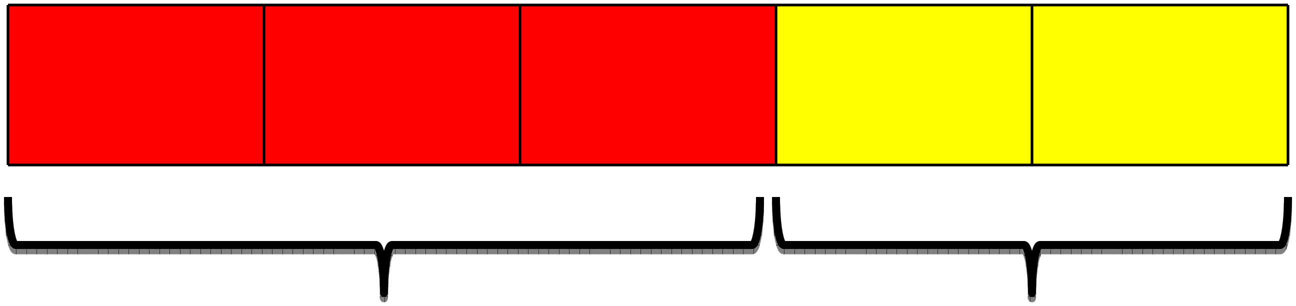
- 183 documents in TimeBank



Rules Development

Experimental Setup

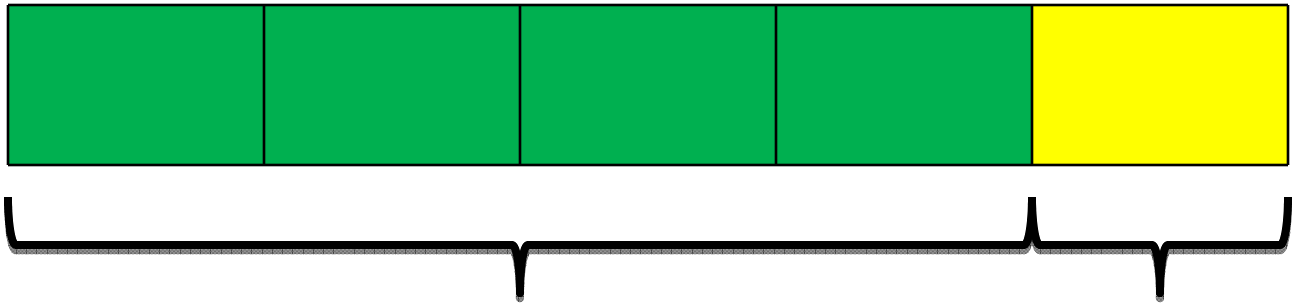
- 183 documents in TimeBank



Rules Development **2 fold cross validation**

Experimental Setup

- 183 documents in TimeBank

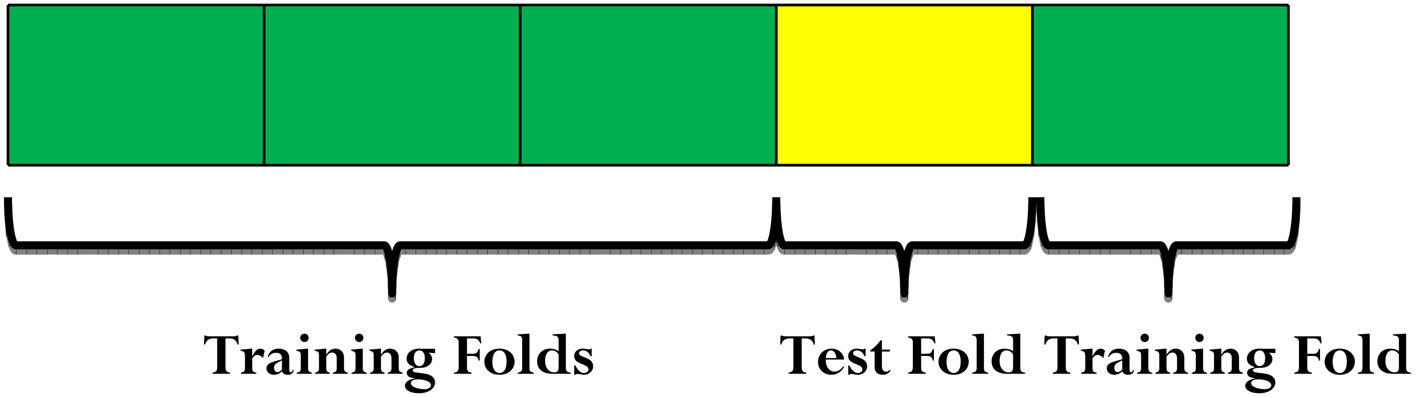


Training Folds

Test Fold

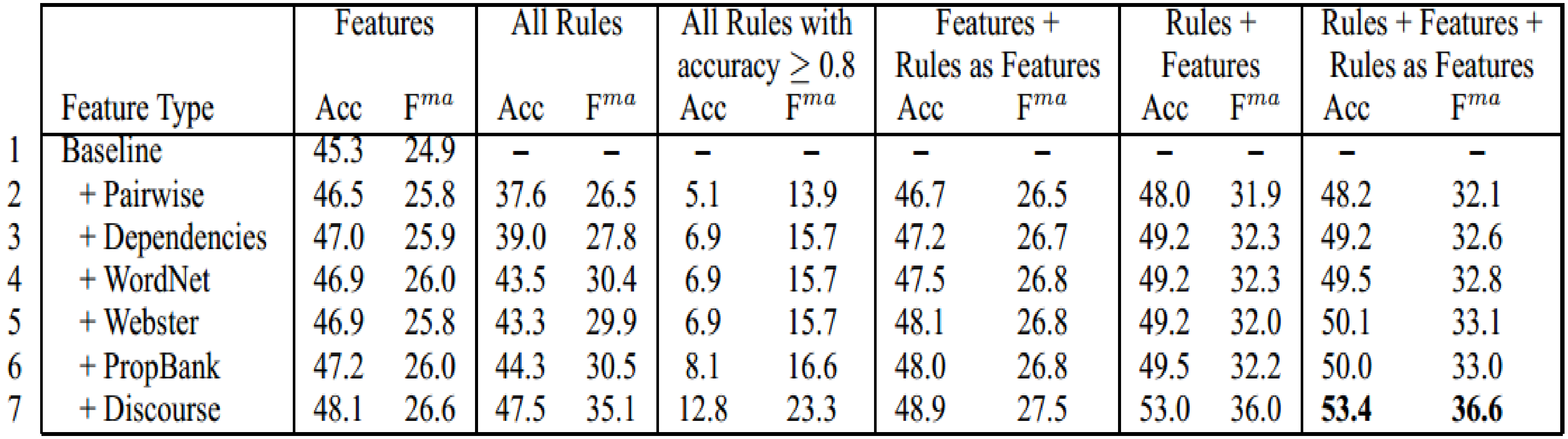
Experimental Setup

- 183 documents in TimeBank



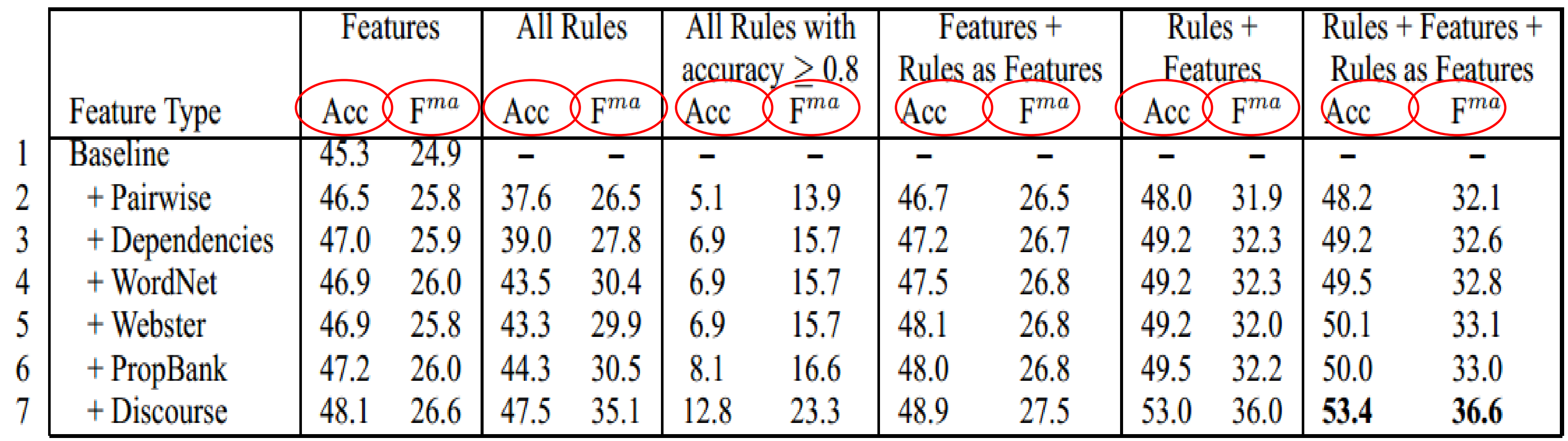
- Evaluation metrics:
  - Accuracy: % of correctly classified instances
  - Macro Fscore = (sum of the f-scores for each of the 14 temporal relation types)/14

Results



		Features		All Rules		All Rules with accuracy ≥ 0.8		Features + Rules as Features		Rules + Features		Rules + Features + Rules as Features	
	Feature Type	Acc	F^{ma}	Acc	F^{ma}	Acc	F^{ma}	Acc	F^{ma}	Acc	F^{ma}	Acc	F^{ma}
1	Baseline	45.3	24.9	–	–	–	–	–	–	–	–	–	–
2	+ Pairwise	46.5	25.8	37.6	26.5	5.1	13.9	46.7	26.5	48.0	31.9	48.2	32.1
3	+ Dependencies	47.0	25.9	39.0	27.8	6.9	15.7	47.2	26.7	49.2	32.3	49.2	32.6
4	+ WordNet	46.9	26.0	43.5	30.4	6.9	15.7	47.5	26.8	49.2	32.3	49.5	32.8
5	+ Webster	46.9	25.8	43.3	29.9	6.9	15.7	48.1	26.8	49.2	32.0	50.1	33.1
6	+ PropBank	47.2	26.0	44.3	30.5	8.1	16.6	48.0	26.8	49.5	32.2	50.0	33.0
7	+ Discourse	48.1	26.6	47.5	35.1	12.8	23.3	48.9	27.5	53.0	36.0	**53.4**	**36.6**

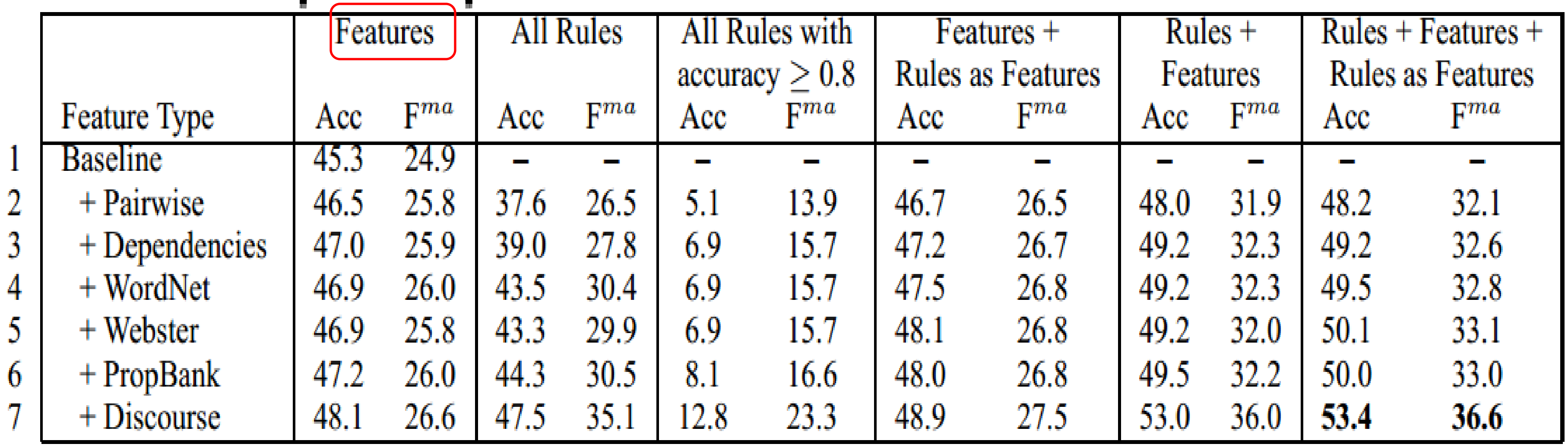
Results



		Features		All Rules		All Rules with accuracy ≥ 0.8		Features + Rules as Features		Rules + Features		Rules + Features + Rules as Features	
	Feature Type	Acc	F^{ma}	Acc	F^{ma}	Acc	F^{ma}	Acc	F^{ma}	Acc	F^{ma}	Acc	F^{ma}
1	Baseline	45.3	24.9	–	–	–	–	–	–	–	–	–	–
2	+ Pairwise	46.5	25.8	37.6	26.5	5.1	13.9	46.7	26.5	48.0	31.9	48.2	32.1
3	+ Dependencies	47.0	25.9	39.0	27.8	6.9	15.7	47.2	26.7	49.2	32.3	49.2	32.6
4	+ WordNet	46.9	26.0	43.5	30.4	6.9	15.7	47.5	26.8	49.2	32.3	49.5	32.8
5	+ Webster	46.9	25.8	43.3	29.9	6.9	15.7	48.1	26.8	49.2	32.0	50.1	33.1
6	+ PropBank	47.2	26.0	44.3	30.5	8.1	16.6	48.0	26.8	49.5	32.2	50.0	33.0
7	+ Discourse	48.1	26.6	47.5	35.1	12.8	23.3	48.9	27.5	53.0	36.0	**53.4**	**36.6**

Results

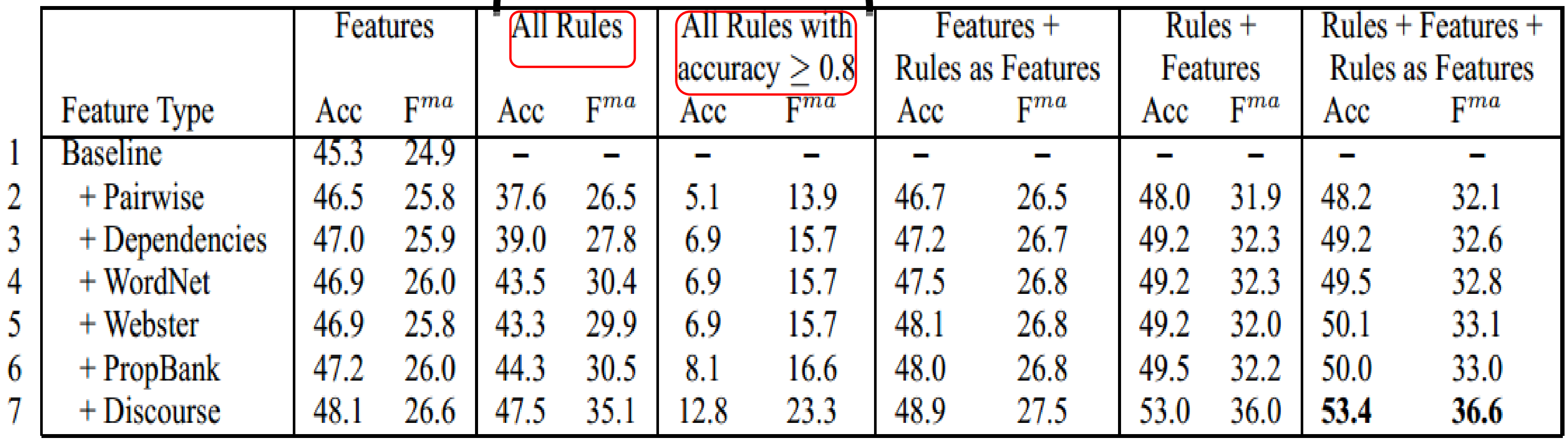
Learning-based System



		Features		All Rules		All Rules with accuracy ≥ 0.8		Features + Rules as Features		Rules + Features		Rules + Features + Rules as Features	
	Feature Type	Acc	F^{ma}	Acc	F^{ma}	Acc	F^{ma}	Acc	F^{ma}	Acc	F^{ma}	Acc	F^{ma}
1	Baseline	45.3	24.9	–	–	–	–	–	–	–	–	–	–
2	+ Pairwise	46.5	25.8	37.6	26.5	5.1	13.9	46.7	26.5	48.0	31.9	48.2	32.1
3	+ Dependencies	47.0	25.9	39.0	27.8	6.9	15.7	47.2	26.7	49.2	32.3	49.2	32.6
4	+ WordNet	46.9	26.0	43.5	30.4	6.9	15.7	47.5	26.8	49.2	32.3	49.5	32.8
5	+ Webster	46.9	25.8	43.3	29.9	6.9	15.7	48.1	26.8	49.2	32.0	50.1	33.1
6	+ PropBank	47.2	26.0	44.3	30.5	8.1	16.6	48.0	26.8	49.5	32.2	50.0	33.0
7	+ Discourse	48.1	26.6	47.5	35.1	12.8	23.3	48.9	27.5	53.0	36.0	**53.4**	**36.6**

Results

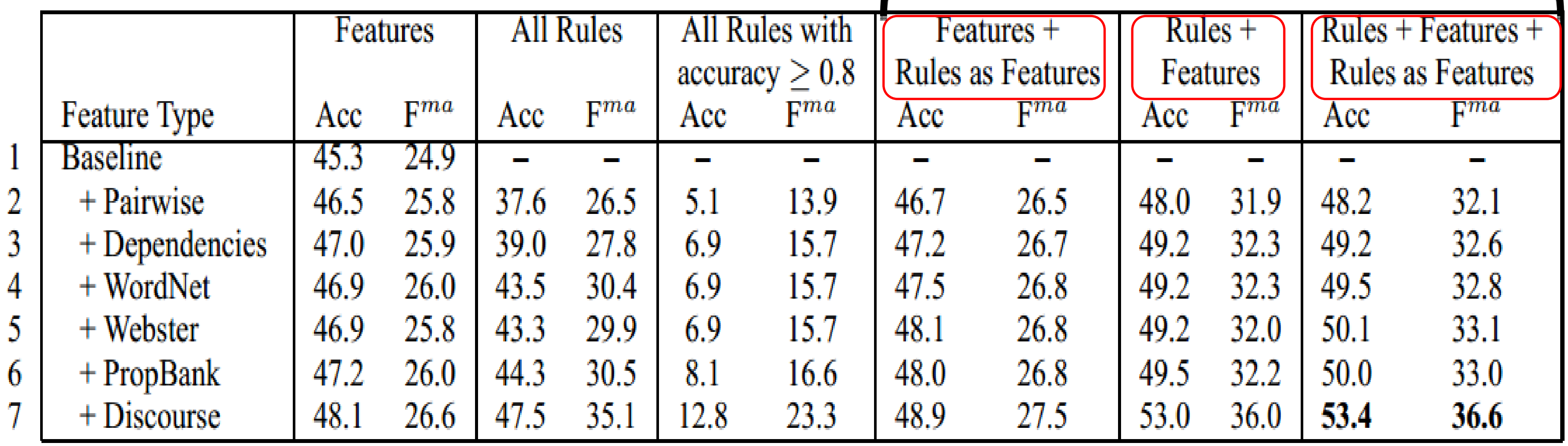
Rule-based Systems



		Features		All Rules		All Rules with accuracy ≥ 0.8		Features + Rules as Features		Rules + Features		Rules + Features + Rules as Features	
	Feature Type	Acc	F^{ma}	Acc	F^{ma}	Acc	F^{ma}	Acc	F^{ma}	Acc	F^{ma}	Acc	F^{ma}
1	Baseline	45.3	24.9	–	–	–	–	–	–	–	–	–	–
2	+ Pairwise	46.5	25.8	37.6	26.5	5.1	13.9	46.7	26.5	48.0	31.9	48.2	32.1
3	+ Dependencies	47.0	25.9	39.0	27.8	6.9	15.7	47.2	26.7	49.2	32.3	49.2	32.6
4	+ WordNet	46.9	26.0	43.5	30.4	6.9	15.7	47.5	26.8	49.2	32.3	49.5	32.8
5	+ Webster	46.9	25.8	43.3	29.9	6.9	15.7	48.1	26.8	49.2	32.0	50.1	33.1
6	+ PropBank	47.2	26.0	44.3	30.5	8.1	16.6	48.0	26.8	49.5	32.2	50.0	33.0
7	+ Discourse	48.1	26.6	47.5	35.1	12.8	23.3	48.9	27.5	53.0	36.0	**53.4**	**36.6**

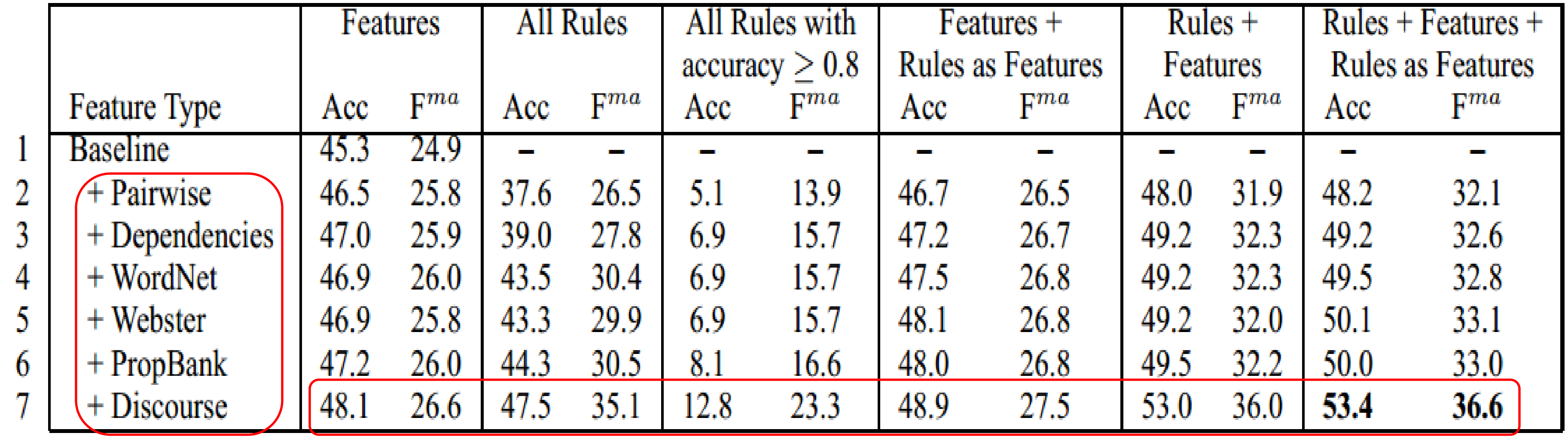
Results

Hybrid Systems



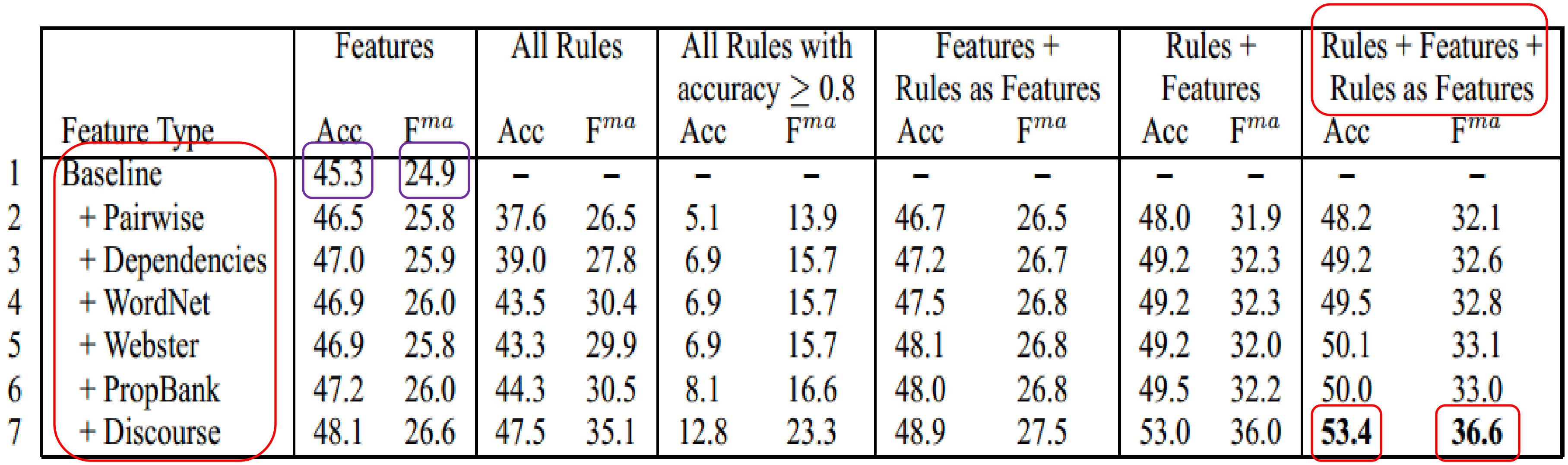
		Features		All Rules		All Rules with accuracy ≥ 0.8		Features + Rules as Features		Rules + Features		Rules + Features + Rules as Features	
	Feature Type	Acc	F^{ma}	Acc	F^{ma}	Acc	F^{ma}	Acc	F^{ma}	Acc	F^{ma}	Acc	F^{ma}
1	Baseline	45.3	24.9	–	–	–	–	–	–	–	–	–	–
2	+ Pairwise	46.5	25.8	37.6	26.5	5.1	13.9	46.7	26.5	48.0	31.9	48.2	32.1
3	+ Dependencies	47.0	25.9	39.0	27.8	6.9	15.7	47.2	26.7	49.2	32.3	49.2	32.6
4	+ WordNet	46.9	26.0	43.5	30.4	6.9	15.7	47.5	26.8	49.2	32.3	49.5	32.8
5	+ Webster	46.9	25.8	43.3	29.9	6.9	15.7	48.1	26.8	49.2	32.0	50.1	33.1
6	+ PropBank	47.2	26.0	44.3	30.5	8.1	16.6	48.0	26.8	49.5	32.2	50.0	33.0
7	+ Discourse	48.1	26.6	47.5	35.1	12.8	23.3	48.9	27.5	53.0	36.0	**53.4**	**36.6**

Results



		Features		All Rules		All Rules with accuracy ≥ 0.8		Features + Rules as Features		Rules + Features		Rules + Features + Rules as Features	
	Feature Type	Acc	F^{ma}	Acc	F^{ma}	Acc	F^{ma}	Acc	F^{ma}	Acc	F^{ma}	Acc	F^{ma}
1	Baseline	45.3	24.9	–	–	–	–	–	–	–	–	–	–
2	+ Pairwise	46.5	25.8	37.6	26.5	5.1	13.9	46.7	26.5	48.0	31.9	48.2	32.1
3	+ Dependencies	47.0	25.9	39.0	27.8	6.9	15.7	47.2	26.7	49.2	32.3	49.2	32.6
4	+ WordNet	46.9	26.0	43.5	30.4	6.9	15.7	47.5	26.8	49.2	32.3	49.5	32.8
5	+ Webster	46.9	25.8	43.3	29.9	6.9	15.7	48.1	26.8	49.2	32.0	50.1	33.1
6	+ PropBank	47.2	26.0	44.3	30.5	8.1	16.6	48.0	26.8	49.5	32.2	50.0	33.0
7	+ Discourse	48.1	26.6	47.5	35.1	12.8	23.3	48.9	27.5	53.0	36.0	**53.4**	**36.6**

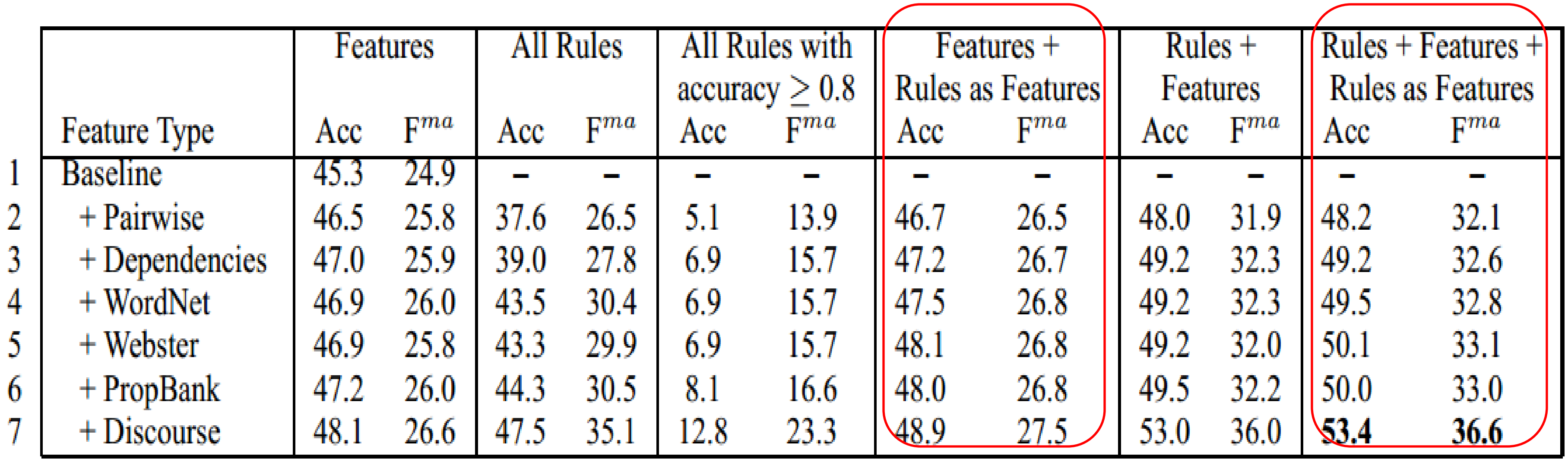
Results



		Features		All Rules		All Rules with accuracy $\geq$ 0.8		Features + Rules as Features		Rules + Features		Rules + Features + Rules as Features	
	Feature Type	Acc	F^{ma}	Acc	F^{ma}	Acc	F^{ma}	Acc	F^{ma}	Acc	F^{ma}	Acc	F^{ma}
1	Baseline	45.3	24.9	–	–	–	–	–	–	–	–	–	–
2	+ Pairwise	46.5	25.8	37.6	26.5	5.1	13.9	46.7	26.5	48.0	31.9	48.2	32.1
3	+ Dependencies	47.0	25.9	39.0	27.8	6.9	15.7	47.2	26.7	49.2	32.3	49.2	32.6
4	+ WordNet	46.9	26.0	43.5	30.4	6.9	15.7	47.5	26.8	49.2	32.3	49.5	32.8
5	+ Webster	46.9	25.8	43.3	29.9	6.9	15.7	48.1	26.8	49.2	32.0	50.1	33.1
6	+ PropBank	47.2	26.0	44.3	30.5	8.1	16.6	48.0	26.8	49.5	32.2	50.0	33.0
7	+ Discourse	48.1	26.6	47.5	35.1	12.8	23.3	48.9	27.5	53.0	36.0	**53.4**	**36.6**

- Improvement over the baseline: 15-16 % relative error reduction
- Goal: Examine
  - impact of different system architectures on performance
  - impact of different feature types on performance

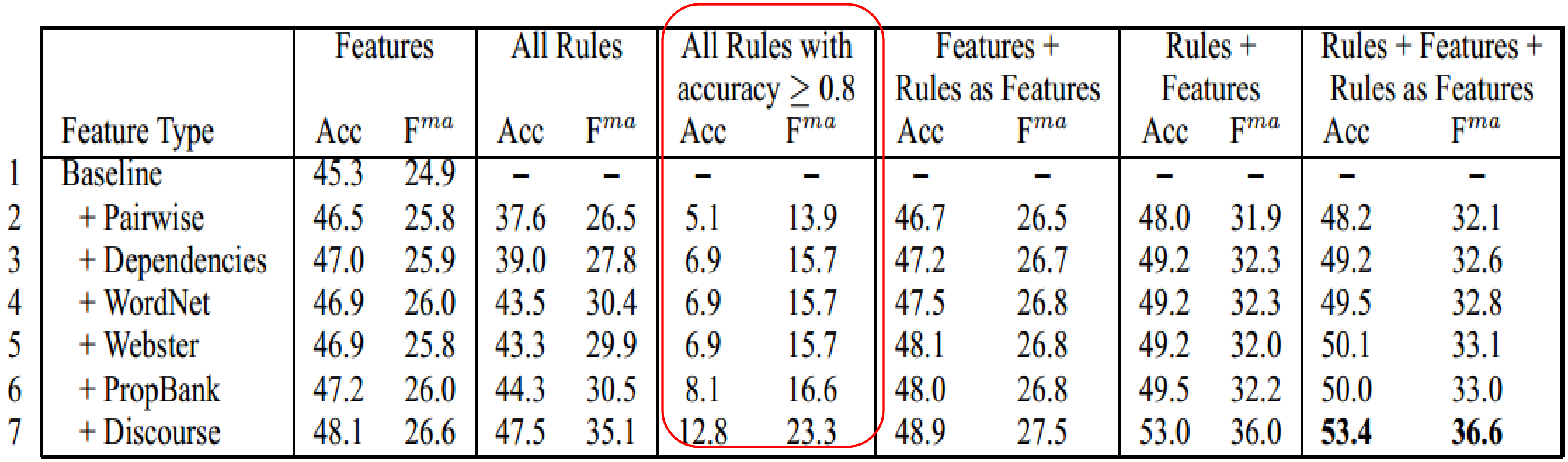
Is the use of Rules useful in the Hybrid Systems?



		Features		All Rules		All Rules with accuracy ≥ 0.8		Features + Rules as Features		Rules + Features		Rules + Features + Rules as Features	
	Feature Type	Acc	F^{ma}	Acc	F^{ma}	Acc	F^{ma}	Acc	F^{ma}	Acc	F^{ma}	Acc	F^{ma}
1	Baseline	45.3	24.9	–	–	–	–	–	–	–	–	–	–
2	+ Pairwise	46.5	25.8	37.6	26.5	5.1	13.9	46.7	26.5	48.0	31.9	48.2	32.1
3	+ Dependencies	47.0	25.9	39.0	27.8	6.9	15.7	47.2	26.7	49.2	32.3	49.2	32.6
4	+ WordNet	46.9	26.0	43.5	30.4	6.9	15.7	47.5	26.8	49.2	32.3	49.5	32.8
5	+ Webster	46.9	25.8	43.3	29.9	6.9	15.7	48.1	26.8	49.2	32.0	50.1	33.1
6	+ PropBank	47.2	26.0	44.3	30.5	8.1	16.6	48.0	26.8	49.5	32.2	50.0	33.0
7	+ Discourse	48.1	26.6	47.5	35.1	12.8	23.3	48.9	27.5	53.0	36.0	**53.4**	**36.6**

- Rules are effective when they are used to make classification decisions prior to the application of the classifier

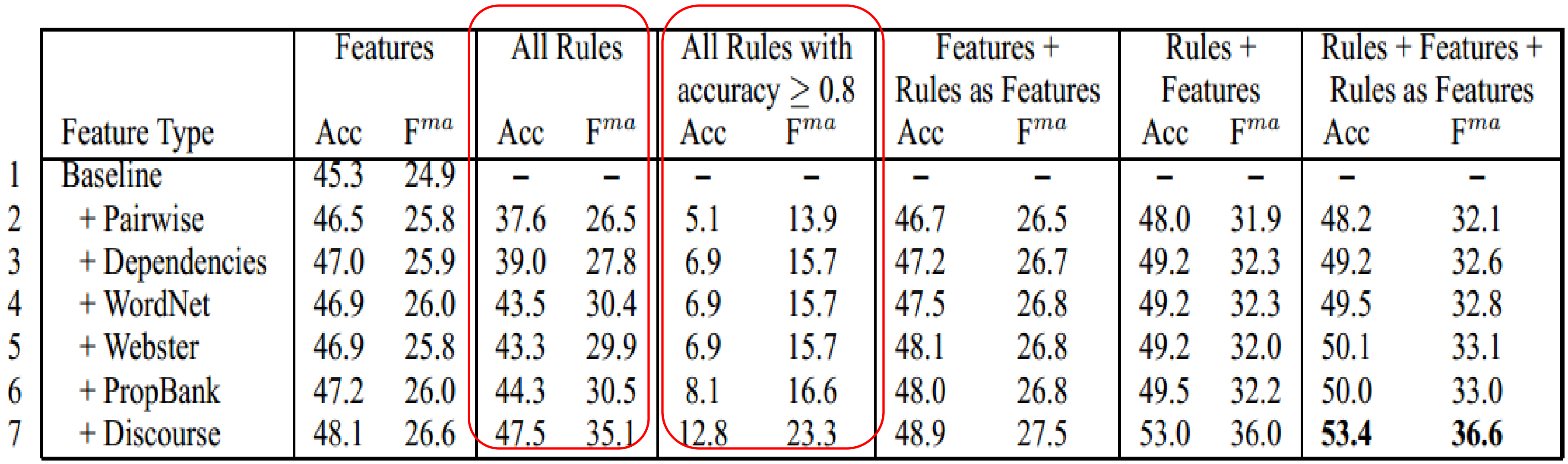
How well do Purely Rule-Based Approaches perform?



		Features		All Rules		All Rules with accuracy ≥ 0.8		Features + Rules as Features		Rules + Features		Rules + Features + Rules as Features	
	Feature Type	Acc	F^{ma}	Acc	F^{ma}	Acc	F^{ma}	Acc	F^{ma}	Acc	F^{ma}	Acc	F^{ma}
1	Baseline	45.3	24.9	–	–	–	–	–	–	–	–	–	–
2	+ Pairwise	46.5	25.8	37.6	26.5	5.1	13.9	46.7	26.5	48.0	31.9	48.2	32.1
3	+ Dependencies	47.0	25.9	39.0	27.8	6.9	15.7	47.2	26.7	49.2	32.3	49.2	32.6
4	+ WordNet	46.9	26.0	43.5	30.4	6.9	15.7	47.5	26.8	49.2	32.3	49.5	32.8
5	+ Webster	46.9	25.8	43.3	29.9	6.9	15.7	48.1	26.8	49.2	32.0	50.1	33.1
6	+ PropBank	47.2	26.0	44.3	30.5	8.1	16.6	48.0	26.8	49.5	32.2	50.0	33.0
7	+ Discourse	48.1	26.6	47.5	35.1	12.8	23.3	48.9	27.5	53.0	36.0	**53.4**	**36.6**

- Rule-based system with only high-accuracy rules has low results owing to low coverage (15.3% recall on test data)

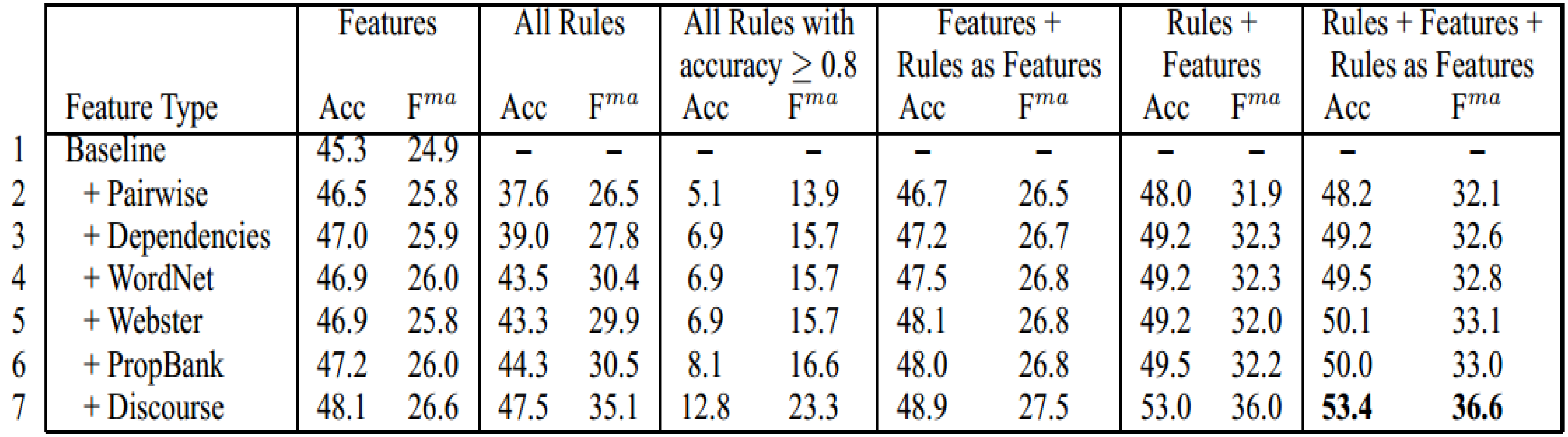
How well do Purely Rule-Based Approaches perform?



		Features		All Rules		All Rules with accuracy ≥ 0.8		Features + Rules as Features		Rules + Features		Rules + Features + Rules as Features	
	Feature Type	Acc	F^{ma}	Acc	F^{ma}	Acc	F^{ma}	Acc	F^{ma}	Acc	F^{ma}	Acc	F^{ma}
1	Baseline	45.3	24.9	–	–	–	–	–	–	–	–	–	–
2	+ Pairwise	46.5	25.8	37.6	26.5	5.1	13.9	46.7	26.5	48.0	31.9	48.2	32.1
3	+ Dependencies	47.0	25.9	39.0	27.8	6.9	15.7	47.2	26.7	49.2	32.3	49.2	32.6
4	+ WordNet	46.9	26.0	43.5	30.4	6.9	15.7	47.5	26.8	49.2	32.3	49.5	32.8
5	+ Webster	46.9	25.8	43.3	29.9	6.9	15.7	48.1	26.8	49.2	32.0	50.1	33.1
6	+ PropBank	47.2	26.0	44.3	30.5	8.1	16.6	48.0	26.8	49.5	32.2	50.0	33.0
7	+ Discourse	48.1	26.6	47.5	35.1	12.8	23.3	48.9	27.5	53.0	36.0	**53.4**	**36.6**

- Rule-based system with only high-accuracy rules has low results owing to low coverage (15.3% recall on test data)
- Using all rules is better than using only high-accuracy rules
- Purely rule-based systems are not as competitive as the hybrid systems

Impact of Feature Types



		Features		All Rules		All Rules with accuracy ≥ 0.8		Features + Rules as Features		Rules + Features		Rules + Features + Rules as Features	
	Feature Type	Acc	F^{ma}	Acc	F^{ma}	Acc	F^{ma}	Acc	F^{ma}	Acc	F^{ma}	Acc	F^{ma}
1	Baseline	45.3	24.9	–	–	–	–	–	–	–	–	–	–
2	+ Pairwise	46.5	25.8	37.6	26.5	5.1	13.9	46.7	26.5	48.0	31.9	48.2	32.1
3	+ Dependencies	47.0	25.9	39.0	27.8	6.9	15.7	47.2	26.7	49.2	32.3	49.2	32.6
4	+ WordNet	46.9	26.0	43.5	30.4	6.9	15.7	47.5	26.8	49.2	32.3	49.5	32.8
5	+ Webster	46.9	25.8	43.3	29.9	6.9	15.7	48.1	26.8	49.2	32.0	50.1	33.1
6	+ PropBank	47.2	26.0	44.3	30.5	8.1	16.6	48.0	26.8	49.5	32.2	50.0	33.0
7	+ Discourse	48.1	26.6	47.5	35.1	12.8	23.3	48.9	27.5	53.0	36.0	**53.4**	**36.6**

- Features that yield significant ($p < 0.05$) improvement:
  - pairwise features, dependency relations, and discourse relations
- Webster features improve accuracy at a lower significance ($p < 0.07$) level.

Conclusion

- Attempted 14 class temporal relation classification
- Proposed a knowledge-rich, hybrid approach
- Best results are achieved by using all feature types and “Rules + Features + Rules as Features” architecture