TEMPORAL RELATION IDENTIFICATION AND CLASSIFICATION IN CLINICAL NOTES

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.

Goal

- Advance the state-of-the-art in temporal relation classification in clinical notes by working on a more complex version of the classification task.
  - Attempt fine-grained 12-class classification as opposed to broader 3-class classification (2012 i2b2 Challenge)

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

Talk Outline

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

Talk Outline

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

Dataset

- i2b2 Temporal Relations Challenge Corpus
  - 310 de-identified discharge summaries annotated with 12 temporal relations.

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

Simultaneous (32.5%)	**Overlap (40.2%)**
Before (11.1%)	**After (4.1%)**
Before_Overlap (3.6%)	**Overlap_After (11.6%)**
During (2.7%)	**During_Inv (4.5%)**
Begins (5.5%)	**Begun_By (1.4%)**
Ends (2.3%)	**Ended_By (3.1%)**

Dataset

- i2b2 Corpus
  - 310 de-identified discharge summaries annotated with 12 temporal relations.

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

Simultaneous (32.5%)		**Overlap (40.2%)**
Before (11.1%)	⟷	**After (4.1%)**
Before_Overlap (3.6%)	⟷	**Overlap_After (11.6%)**
During (2.7%)	⟷	**During_Inv (4.5%)**
Begins (5.5%)	⟷	**Begun_By (1.4%)**
Ends (2.3%)	⟷	**Ended_By (3.1%)**

Dataset

- i2b2 Clinical Temporal Relations Challenge Corpus (i2b2 corpus) [Sun et al., 2013]
  - 310 de-identified discharge summaries annotated with 12 temporal relations.

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

Simultaneous (32.5%)		Overlap (40.2%)
Before (11.1%)	⟷	After (4.1%)
Before_Overlap (3.6%)	⟷	Overlap_After (11.6%)
During (2.7%)	⟷	During_Inv (4.5%)
Begins (5.5%)	⟷	Begun_By (1.4%)
Ends (2.3%)	⟷	Ended_By (3.1%)

Talk Outline

- 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 12 temporal relation types
- Conditions to form a training instance
  - entity1 precedes entity2 in the associated text
  - (entity1, entity2) belongs to one of the 12 temporal relation types

Learning-based Baseline Temporal Relation Classifier

- **67 features**
  1. Lexical (17) – based on word strings from the entity and its context.
  2. Grammatical (33) – based on grammatical syntax including POS and phrase information
  3. Entity attributes (8) – encode type, modality, and polarity of event, or type of time expression
  4. Semantic (4) – based on related temporal arguments, WordNet synsets, and VerbOcean relations
  5. Distance (2)
  6. Section creation time related (3)

Learning-based Baseline Temporal Relation Classifier

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

- Specialized Classifiers
  - Following Tang et al., 2012 we train four specialized classifiers rather than one
    - **Intra-sentence event-event classifier**
    - **Intra-sentence event-time classifier**
    - **Inter-sentence event-event classifier**
    - **Inter-sentence coreferent event classifier**

Talk Outline

- 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. Type and modality of entity1 with type and modality of entity2

Event of type Treatment and modality Factual

- E.g.: Patient was given supplemental oxygen for shortness of breath.

Event of type Problem and modality Factual

- supplemental oxygen OVERLAP_AFTER shortness of breath
- Feature value: $TREATMENT_1$-$FACTUAL_1$-$PROBLEM_2$-$FACTUAL_2$

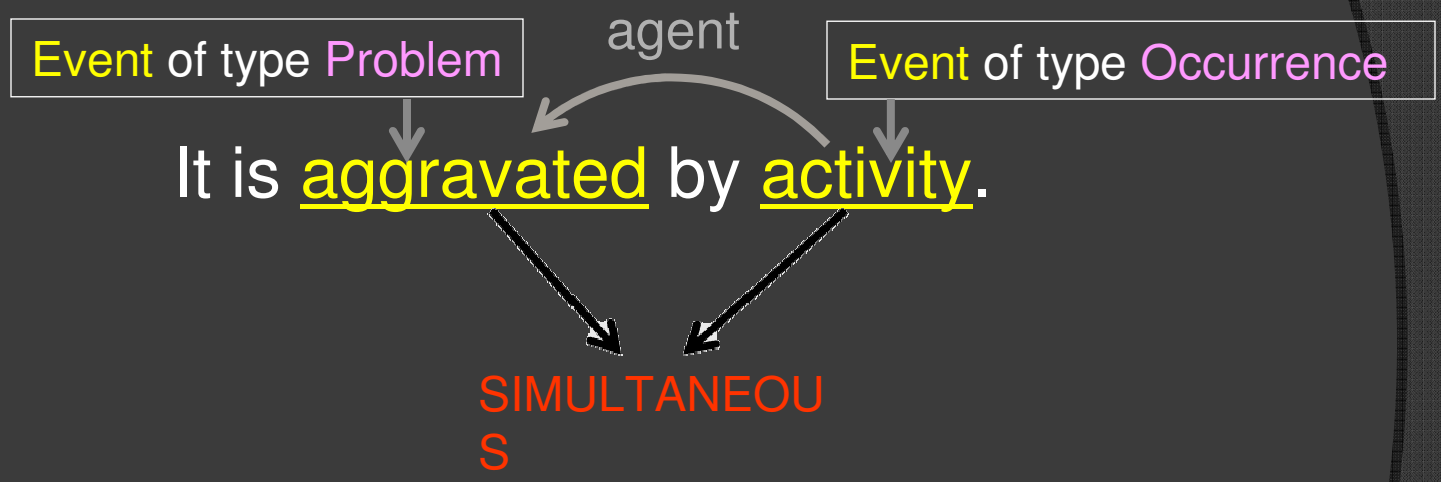
Pairwise Features

2. Entity head word pairs
3. Prepositional lexeme pairs
4. Preposition trace feature
5. 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?



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

Dependency Relation Features

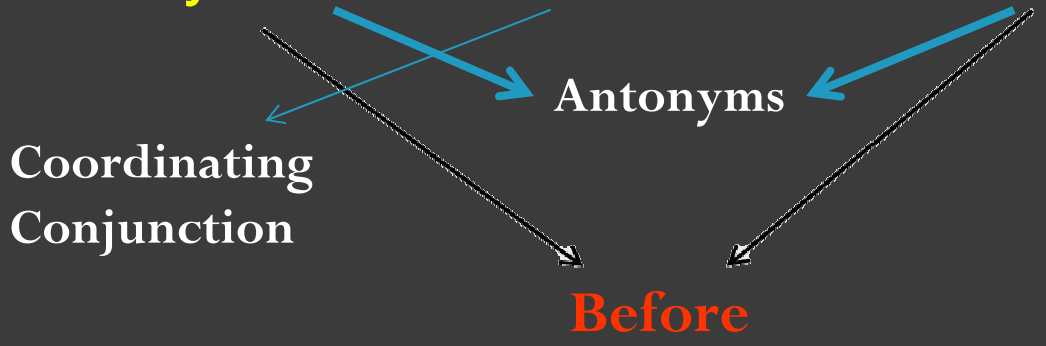
- For each of the 25 dependency relation types produced by the Stanford parser we form binary features:
  - Is entity1/entity2 the governor in the relation?
  - Is entity1/entity2 the dependent in the relation?

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?

Her amylase was mildly elevated but has been down since then.



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) ∈ t?
    - is (event2, event1) ∈ t?

WordNet Relation Features

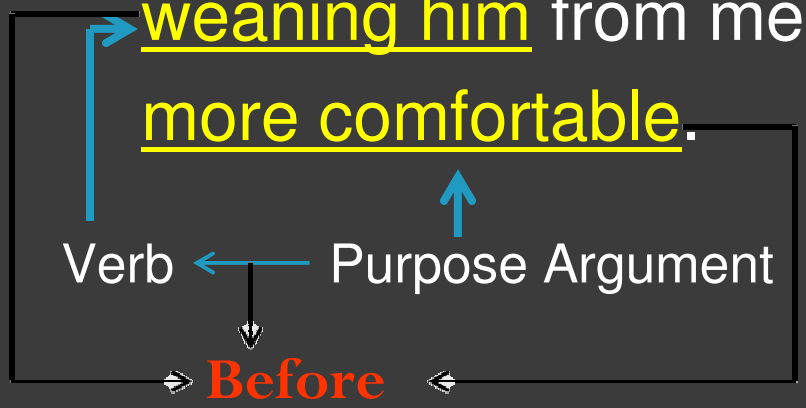
- 4 types of WordNet semantic relations:
  - hypernym, hyponym, troponym, and similar
- 8 binary features:
  - for each type of semantic relation:
    - is (event1, event2) ∈ t?
    - is (event2, event1) ∈ t?

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?

Discussion should occur with the family about weaning him from medications to make him more comfortable.



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

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

What are discourse relations?

At operation, there was no gross adenopathy, and it was felt that the tumor was completely excised.

The patient thereafter had a benign convalescence.

- Explicit Relation: Asynchronous

What are discourse relations?

ARG_1 { At operation, there was no gross adenopathy, and it was felt that the tumor was completely excised.

ARG_2 { The patient thereafter had a benign convalescence. }

- Explicit Relation: Asynchronous

Why are Discourse Relations useful for Temporal Relation Classification?

Event of type Treatment

At operation, there was no gross adenopathy, and it was felt that the tumor was completely excised.

The patient thereafter had a benign convalescence.

Event of type Problem

- Explicit Relation: Asynchronous

Why are Discourse Relations useful for Temporal Relation Classification?

At operation, there was no gross adenopathy, and it was felt that the tumor was completely excised.

The patient thereafter had a benign convalescence.

- Intuitively, a treatment event within a discourse unit happens before an occurrence event contained within a separate asynchronous discourse unit.

Why are Discourse Relations useful for Temporal Relation Classification?

At operation, there was no gross adenopathy, and it was felt that the tumor was completely excised. The patient thereafter had a benign convalescence.

Discourse relations can potentially be exploited to discover both inter-sentential and intra-sentential temporal relations

Why are Discourse Relations useful for Temporal Relation Classification?

At operation, there was no gross adenopathy, and it was felt that the tumor was completely excised. The patient thereafter had a benign convalescence.

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

Discourse Relation Features

- 12 types of discourse relations (extracted 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?
- Same 48 features based on implicit relations

Talk Outline

- 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 &&
        entity1.class=TREATMENT &&
        entity2.class=PROBLEM &&
        explicitDiscourseArg1ConditionArg2(entity1, entity2)
then infer relation=AFTER;
```

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

- 2 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 75% accurate. If none of the rules is applicable, we classify it using the classifier employed in method 1.

Talk Outline

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

Experimental Setup

- i2b2 corpus:
  - 190 training documents
  - 120 test documents

Experimental Setup

- Evaluation metric:
  - Micro Fscore = harmonic mean of single precision and recall computed over all classes.

Results

	Features	All Rules	All Rules with accuracy >= 0.75	Features + Rules as Features	Rules + Features + Rules as Features
Feature Type	micro F	micro F	micro F	micro F	micro F
Baseline	55.3				
+ Pairwise	55.5	37.6	14.5	57.2	57.8
+ Dependencies	55.5	40.0	16.2	57.4	58.1
+ WordNet	55.6	40.0	16.2	57.2	57.9
+ Webster	55.8	40.0	16.2	57.3	58.0
+ PropBank	55.8	45.4	21.3	57.6	59.7
+ Discourse	56.2	47.3	24.0	57.9	61.1

Results

Learning-based System

	Features	All Rules	All Rules with accuracy >= 0.75	Features + Rules as Features	Rules + Features + Rules as Features
Feature Type	micro F	micro F	micro F	micro F	micro F
Baseline	55.3				
+ Pairwise	55.5	37.6	14.5	57.2	57.8
+ Dependencies	55.5	40.0	16.2	57.4	58.1
+ WordNet	55.6	40.0	16.2	57.2	57.9
+ Webster	55.8	40.0	16.2	57.3	58.0
+ PropBank	55.8	45.4	21.3	57.6	59.7
+ Discourse	56.2	47.3	24.0	57.9	61.1

Results

Rule-based Systems

	Features	All Rules	All Rules with accuracy >= 0.75	Features + Rules as Features	Rules + Features + Rules as Features
Feature Type	micro F	micro F	micro F	micro F	micro F
Baseline	55.3				
+ Pairwise	55.5	37.6	14.5	57.2	57.8
+ Dependencies	55.5	40.0	16.2	57.4	58.1
+ WordNet	55.6	40.0	16.2	57.2	57.9
+ Webster	55.8	40.0	16.2	57.3	58.0
+ PropBank	55.8	45.4	21.3	57.6	59.7
+ Discourse	56.2	47.3	24.0	57.9	61.1

Results

Hybrid Systems

	Features	All Rules	All Rules with accuracy >= 0.75	Features + Rules as Features	Rules + Features + Rules as Features
Feature Type	micro F	micro F	micro F	micro F	micro F
Baseline	55.3				
+ Pairwise	55.5	37.6	14.5	57.2	57.8
+ Dependencies	55.5	40.0	16.2	57.4	58.1
+ WordNet	55.6	40.0	16.2	57.2	57.9
+ Webster	55.8	40.0	16.2	57.3	58.0
+ PropBank	55.8	45.4	21.3	57.6	59.7
+ Discourse	56.2	47.3	24.0	57.9	61.1

Results

	Features	All Rules	All Rules with accuracy >= 0.75	Features + Rules as Features	Rules + Features + Rules as Features
Feature Type	micro F	micro F	micro F	micro F	micro F
Baseline	55.3				
+ Pairwise	55.5	37.6	14.5	57.2	57.8
+ Dependencies	55.5	40.0	16.2	57.4	58.1
+ WordNet	55.6	40.0	16.2	57.2	57.9
+ Webster	55.8	40.0	16.2	57.3	58.0
+ PropBank	55.8	45.4	21.3	57.6	59.7
+ Discourse	56.2	47.3	24.0	57.9	61.1

Results

	Features	All Rules	All Rules with accuracy >= 0.75	Features + Rules as Features	Rules + Features + Rules as Features
Feature Type	micro F	micro F	micro F	micro F	micro F
Baseline	55.3				
+ Pairwise	55.5	37.6	14.5	57.2	57.8
+ Dependencies	55.5	40.0	16.2	57.4	58.1
+ WordNet	55.6	40.0	16.2	57.2	57.9
+ Webster	55.8	40.0	16.2	57.3	58.0
+ PropBank	55.8	45.4	21.3	57.6	59.7
+ Discourse	56.2	47.3	24.0	57.9	61.1

- Improvement over the baseline: 15% relative error reduction

Which of the non-hybrid systems is the strongest?

Non-Hybrid Systems

	Features	All Rules	All Rules with accuracy >= 0.75	Features + Rules as Features	Rules + Features + Rules as Features
Feature Type	micro F	micro F	micro F	micro F	micro F
Baseline	55.3				
+ Pairwise	55.5	37.6	14.5	57.2	57.8
+ Dependencies	55.5	40.0	16.2	57.4	58.1
+ WordNet	55.6	40.0	16.2	57.2	57.9
+ Webster	55.8	40.0	16.2	57.3	58.0
+ PropBank	55.8	45.4	21.3	57.6	59.7
+ Discourse	56.2	47.3	24.0	57.9	61.1

Which of the non-hybrid systems is the strongest?

Non-Hybrid Systems

	Features	All Rules	All Rules with accuracy >= 0.75	Features + Rules as Features	Rules + Features + Rules as Features
Feature Type	micro F	micro F	micro F	micro F	micro F
Baseline	55.3				
+ Pairwise	55.5	37.6	14.5	57.2	57.8
+ Dependencies	55.5	40.0	16.2	57.4	58.1
+ WordNet	55.6	40.0	16.2	57.2	57.9
+ Webster	55.8	40.0	16.2	57.3	58.0
+ PropBank	55.8	45.4	21.3	57.6	59.7
+ Discourse	56.2	47.3	24.0	57.9	61.1

Are rules when used as features helpful?

	Features	All Rules	All Rules with accuracy >= 0.75	Features + Rules as Features	Rules + Features + Rules as Features
Feature Type	micro F	micro F	micro F	micro F	micro F
Baseline	55.3				
+ Pairwise	55.5	37.6	14.5	57.2	57.8
+ Dependencies	55.5	40.0	16.2	57.4	58.1
+ WordNet	55.6	40.0	16.2	57.2	57.9
+ Webster	55.8	40.0	16.2	57.3	58.0
+ PropBank	55.8	45.4	21.3	57.6	59.7
+ Discourse	56.2	47.3	24.0	57.9	61.1

Impact of feature types

	Features	All Rules	All Rules with accuracy >= 0.75	Features + Rules as Features	Rules + Features + Rules as Features
Feature Type	micro F	micro F	micro F	micro F	micro F
Baseline	55.3				
+ Pairwise	55.5	37.6	14.5	57.2	57.8
+ Dependencies	55.5	40.0	16.2	57.4	58.1
+ WordNet	55.6	40.0	16.2	57.2	57.9
+ Webster	55.8	40.0	16.2	57.3	58.0
+ PropBank	55.8	45.4	21.3	57.6	59.7
+ Discourse	56.2	47.3	24.0	57.9	61.1

- Features that yield significant improvement
  - pairwise features, predicate-argument relations, and discourse relations

Impact of feature types

	Features	All Rules	All Rules with accuracy >= 0.75	Features + Rules as Features	Rules + Features + Rules as Features
Feature Type	micro F	micro F	micro F	micro F	micro F
Baseline	55.3				
+ Pairwise	55.5	37.6	14.5	57.2	57.8
+ Dependencies	55.5	40.0	16.2	57.4	58.1
+ WordNet	55.6	40.0	16.2	57.2	57.9
+ Webster	55.8	40.0	16.2	57.3	58.0
+ PropBank	55.8	45.4	21.3	57.6	59.7
+ Discourse	56.2	47.3	24.0	57.9	61.1

- Features that are not useful
  - dependencies, wordnet, and webster features

Results

- So far… classified entity pairs that are known to belong to one of the 12 temporal relations.

- Next… a more challenging evaluation setting
  - classify entity pairs that may or may not have a temporal relation
    - Need to deal with an additional class: the “no temporal relation” class.

Results for the challenging setting

	Features	All Rules	All Rules with accuracy >= 0.75	Features + Rules as Features	Rules + Features + Rules as Features
Feature Type	micro F	micro F	micro F	micro F	micro F
Baseline	26.0				
+ Pairwise	26.5	14.8	6.8	26.6	27.5
+ Dependencies	26.5	17.1	9.9	26.7	27.7
+ WordNet	26.5	17.2	9.9	26.6	27.6
+ Webster	26.5	17.2	9.9	26.7	27.6
+ PropBank	26.5	21.2	15.4	26.8	29.1
+ Discourse	26.6	21.9	18.8	26.8	30.0

- As expected these results are lower than those with known temporal relations.

Results for the challenging setting

	Features	All Rules	All Rules with accuracy >= 0.75	Features + Rules as Features	Rules + Features + Rules as Features
Feature Type	micro F	micro F	micro F	micro F	micro F
Baseline	26.0				
+ Pairwise	26.5	14.8	6.8	26.6	27.5
+ Dependencies	26.5	17.1	9.9	26.7	27.7
+ WordNet	26.5	17.2	9.9	26.6	27.6
+ Webster	26.5	17.2	9.9	26.7	27.6
+ PropBank	26.5	21.2	15.4	26.8	29.1
+ Discourse	26.6	21.9	18.8	26.8	30.0

- The impact of the system architectures and feature types in this challenging setting is the same as before.

Results for the challenging setting

	Features	All Rules	All Rules with accuracy >= 0.75	Features + Rules as Features	Rules + Features + Rules as Features
Feature Type	micro F	micro F	micro F	micro F	micro F
Baseline	26.0				
+ Pairwise	26.5	14.8	6.8	26.6	27.5
+ Dependencies	26.5	17.1	9.9	26.7	27.7
+ WordNet	26.5	17.2	9.9	26.6	27.6
+ Webster	26.5	17.2	9.9	26.7	27.6
+ PropBank	26.5	21.2	15.4	26.8	29.1
+ Discourse	26.6	21.9	18.8	26.8	30.0

- Improvement over the baseline: 6% relative error reduction

Conclusion

- Attempted 12 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