A Structured Learning Approach to Temporal Relation Extraction

AntNLP -- Changzhi Sun

changzhisun@stu.ecnu.edu.cn

A Structured Learning Approach to Temporal Relation Extraction

Qiang Ning[1] and **Zhili Feng**[2] and **Dan Roth**[1,2]
[1]Department of Electrical and Computer Engineering
[2]Department of Computer Science
University of Illinois, Urbana, IL 61801
{qning2,zfeng6,danr}@illinois.edu

Outline

1. Introduction
2. Background
3. Approach
4. Experiments
5. Conclusion

Introduction

- Identifying temporal relations between events is an essential step towards natural language understanding
  - Time-slot filling, storyline construction, clinical narratives processing, temporal question answering
- TempEval (TE) workshops
  - time expression extraction (timex) and normalization
    - `Heidel-Time`, `SUTime`, `IllinoisTime` ...
    - end-to-end `F1` scores being around `80%`
  - temporal relation (TLINKs) extraction
    - `F1` scores of around `35%` in the TE workshops

Temporal Relation Extraction

- Generating a `directed temporal graph`
  - `nodes` temporal entities (i.e., events or timexes)
  - `edges` the TLINSKs between them
  - TLINK annotation is quadratic in `#node`
  - overwhelming fraction of the TLINKs are missing in human annotation
- Three types
  - E-E TLINKs
  - T-T TLINKs
  - E-T TLINKs

Ex1 . . . *tons of earth* ***cascaded*** *down a hillside,* ***ripping*** *two houses from their foundations. No one was* ***hurt****, but firefighters* ***ordered*** *the evacuation of nearby homes and said they'll* ***monitor*** *the shifting ground.* . . .

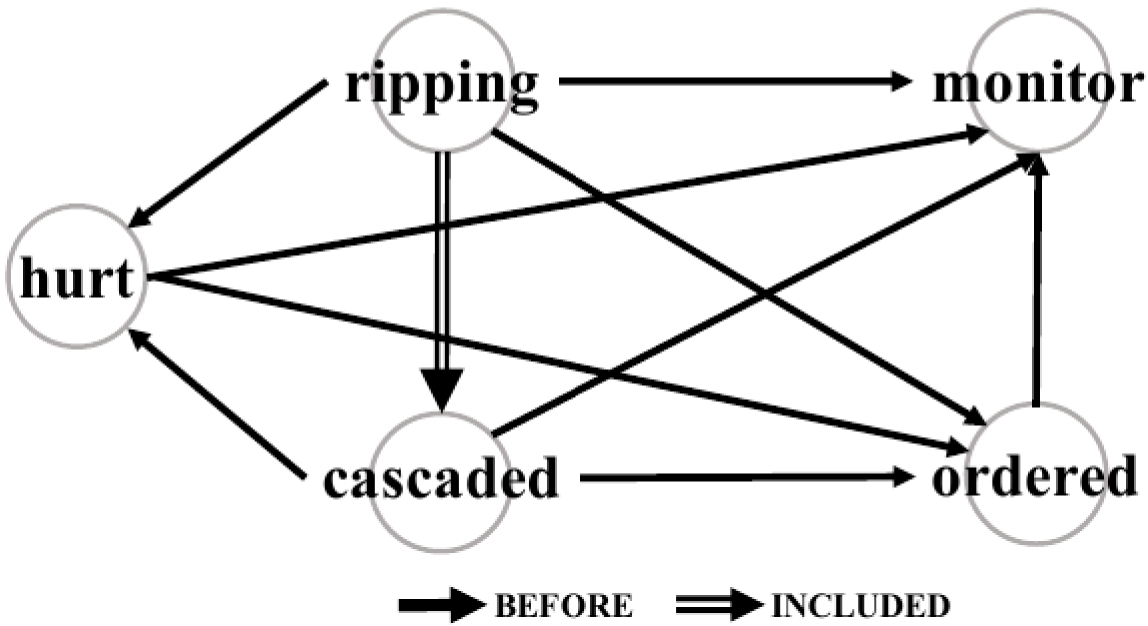

Figure 1: The desired event temporal graph for Ex1. Reverse TLINKs such as ***hurt*** is *after* ***ripping*** are omitted for simplicity.

Temporal Graph

- Symmetry
  - A `before` B $\Rightarrow$ B `after` A
- Transitivity
  - A `before` B and B `before` C $\Rightarrow$ A `before` C
- Making nodes highly interrelated
- The inter-annotator agreement is usually about 50%-60%

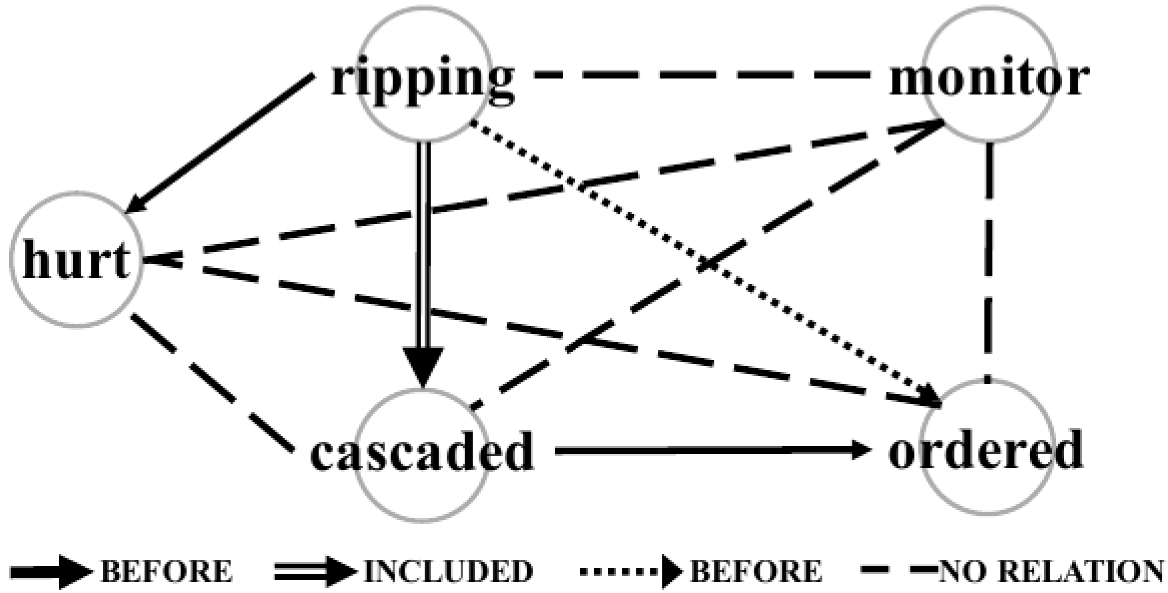

Figure 2: The human-annotation for Ex1 provided in TE3, where many TLINKs are missing due to the annotation difficulty. Solid lines: original human annotations. Dotted lines: TLINKs inferred from solid lines. Dashed lines: missing relations.

Related Work

- Local
  - make pairwise decisions
  - globally inconsistent
- Local + Inference
  - ILP imposes `global constraints` in the `inference` phase
- Global considerations are necessary in the learning phase
  - structured learning approach
  - local models are updated based on feedback from `global inferences`
  - `semi-supervised` method

Temporal Relation Types

- 13 relation types
  - `vague` or `none` is also included when a TLINK is not clear or missing
- reduced set of relation types
  - non-uniform distribution of all the relation types
    - `immediately_before` to `before`
    - `immediately_after` to `after`
  - Due to the ambiguity in natural language
    - `before` to `immediately_before`
  - `before`, `after`, `includes`, `is_included`, `equal`, `vague`

Quality of A Temporal Graph

- The temporal awareness

$$P = \frac{|G_{sys}^{-} \bigcap G_{true}^{+}|}{|G_{sys}^{-}|}$$

$$R = \frac{|G_{true}^{-} \bigcap G_{sys}^{+}|}{|G_{true}^{-}|}$$

- G^{+} is the closure of graph G
- G^{-} is the reduction of G
- `vague` are usually considered as non-existing TLINKs and are not counted during evaluation

Precision = (# of system temporal relations that can be verified from reference annotation temporal closure graph / # of temporal relations in system output)

Recall = (# of reference annotation temporal relations that can be verified from system output's temporal closure graph / # of temporal relations in reference annotation)

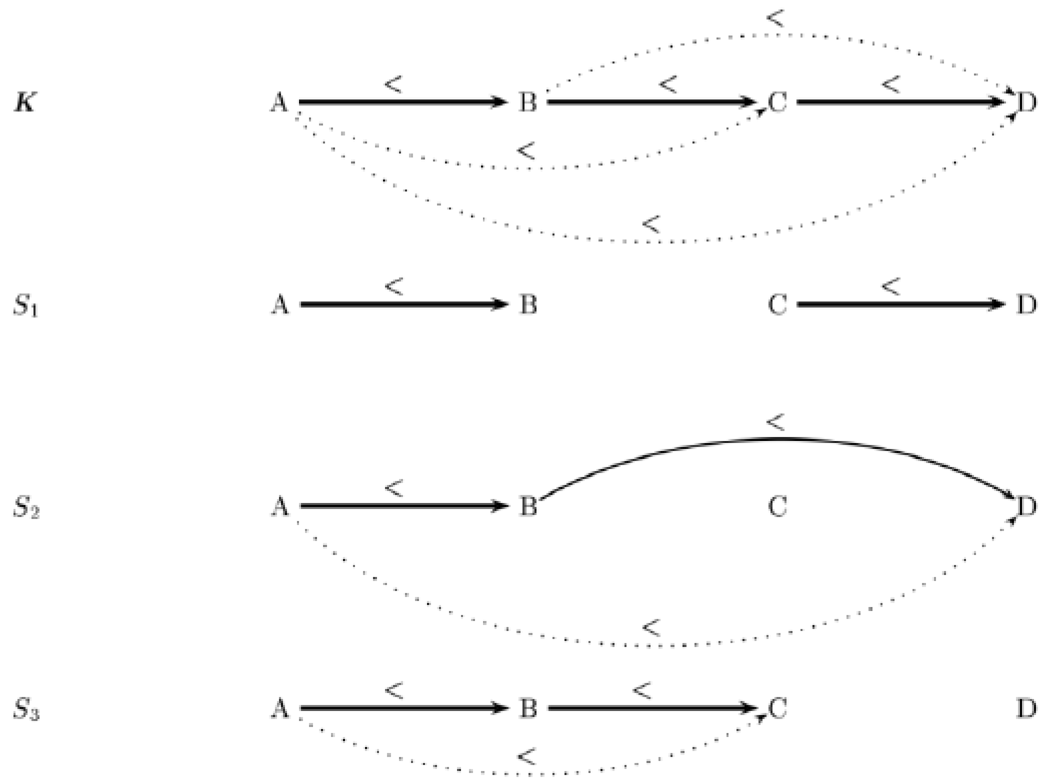

Figure 1: Examples of temporal graphs and relations

System	Precision	Recall	Fscore
S_1	2/2=1	2/3=0.66	0.8
S_2	2/2=1	1/3=0.33	0.5
S_3	2/2=1	2/3=0.66	0.8

Table 1: Precision, recall and fscore for systems in Figure 1 according to our evaluation metric

A Structured Training Approach

- Inference Based Training
- train local classifiers with feedback that accounts for other relations
- performing global inference in each round of the learning

Inference

- n pairs of events
- $\phi_i \in \mathcal{X} \subseteq \mathbb{R}^d$
- $y_i \in \mathcal{Y}$ for the i-th pair of events, $i = 1, 2, ..., n$
    - $\mathcal{Y} = \{r_j\}_{j=1}^{6}$
- $\mathbf{x} = \{\phi_1, ..., \phi_n\} \in \mathcal{X}^n$, $\mathbf{y} = \{y_1, ..., y_n\} \in \mathcal{Y}^n$
- weight vector $\mathbf{w}_r$ of a linear classifier trainded for relation $r \in \mathcal{Y}$ (using the one-vs-all scheme)

Inference

$$\hat{\mathbf{y}} = \arg \max_{\mathbf{y} \in \mathcal{C}(\mathcal{Y}^n)} f(\mathbf{x}, \mathbf{y}),$$

$$f(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} f_{y_i}(\phi_i) = \sum_{i=1}^{n} \frac{e^{\mathbf{w}_{y_i}^T \phi_i}}{\sum_{r \in \mathcal{Y}} e^{\mathbf{w}_r^T \phi_i}}.$$

- $\mathcal{C}(\mathcal{Y}^n) \subseteq \mathcal{Y}^n$ constrains the temporal graph to be symmetrically and transitively consistent
- $f(\mathbf{x}, \mathbf{y})$ is the scoring function
- $f_{y_i}(\phi_i)$ is the prob of the i-th event pair having relation y_i

Inference

- $\mathcal{C}(\mathcal{Y}^n) = \mathcal{Y}^n$
  - Local method
- $\mathcal{C}(\mathcal{Y}^n) \neq \mathcal{Y}^n$
  - as an ILP problem

- $\mathcal{I}_r(ij) \in \{0, 1\}$ be the indicator function of relation r for event event i and event j
- $f_r(ij) \in [0, 1]$ be the corresponding soft-max score

$$\hat{\mathcal{I}} = \underset{\mathcal{I}}{\text{argmax}} \sum_{ij \in \mathcal{E}} \sum_{r \in \mathcal{Y}} f_r(ij) \mathcal{I}_r(ij)$$

$$\text{s.t.} \quad \underbrace{\Sigma_r \mathcal{I}_r(ij) = 1}_{\text{(uniqueness)}}, \ \underbrace{\mathcal{I}_r(ij) = \mathcal{I}_{\bar{r}}(ji)}_{\text{(symmetry)}},$$

$$\underbrace{\mathcal{I}_{r_1}(ij) + \mathcal{I}_{r_2}(jk) - \Sigma_{m=1}^{N} \mathcal{I}_{r_3^m}(ik) \leq 1}_{\text{(transitivity)}},$$

- $\mathcal{E} = \{ij | \text{sentence dist}(i, j) \leq 1\}$
- $\bar{r}$ is the reverse of r
- N is the number of possible relations for r_3 when r_1 and r_2 are true

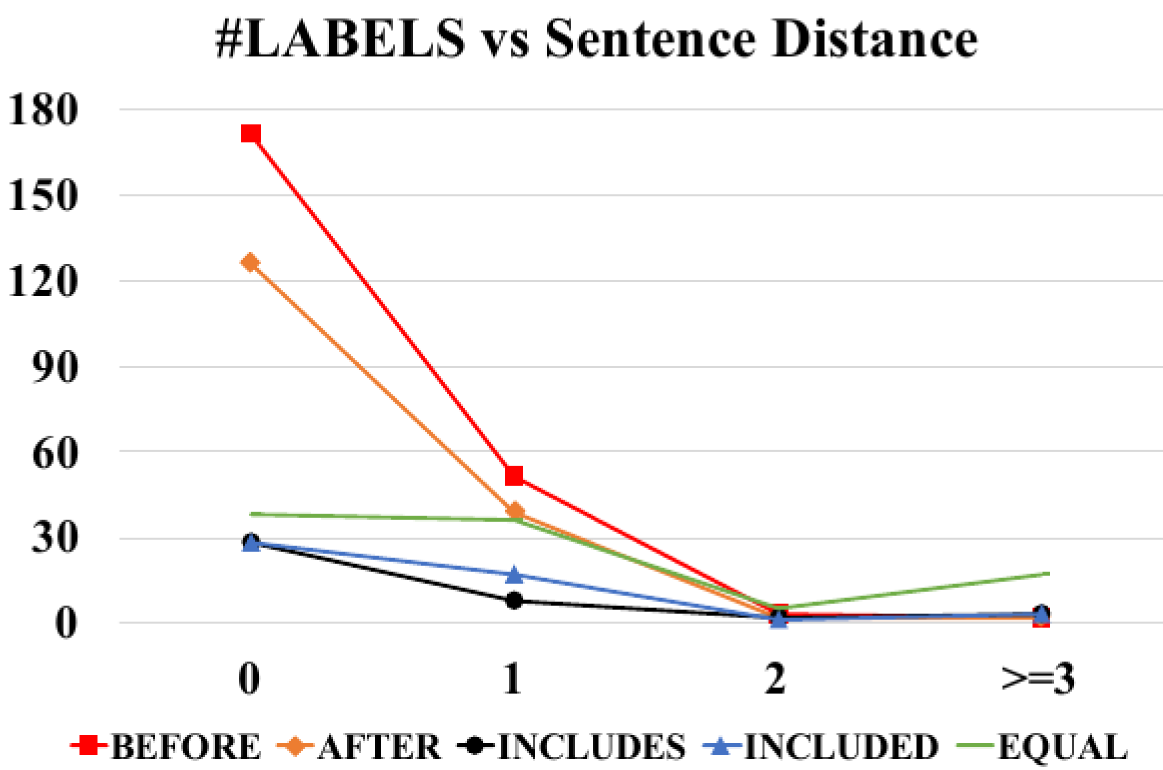

Figure 3: #TLINKs vs sentence distance on the TE3 Platinum dataset. The tail of *equal* is due to event coreference and beyond our focus.

- `pre-filtering` strategy to balance the trade-off between confidence in $f_r(ij)$ and global constraints

- Previous transitivity constraints were formulated as

$$\mathcal{I}_{r_1}(ij) + \mathcal{I}_{r_2}(jk) - \mathcal{I}_{r_3}(ik) \leq 1$$

  - r_1 and r_2 determine a single r_3

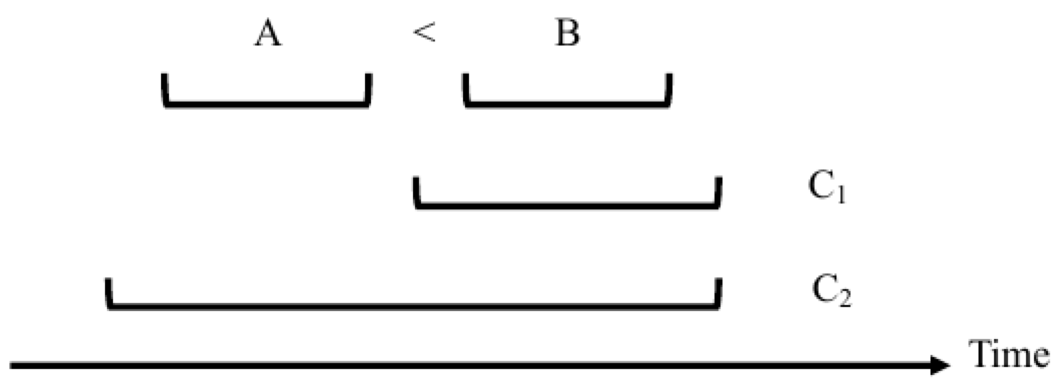

Figure 4: When A *is before* B and B *is_included in* C, A can either be *before* C_1 or *is_included in* C_2. We propose to incorporate this via the transitivity constraints for Eq. (2).

Learning

Algorithm 1: Structured perceptron algorithm for temporal relations

Input: Training set $\mathcal{L} = \{\mathbf{x}_k, \mathbf{y}_k\}_{k=1}^{K}$,
learning rate λ

Perform graph closure on each $\mathbf{y}_k$
Initialize $\mathbf{w}_r = \mathbf{0}, \forall r \in \mathcal{Y}$
while *convergence criteria not satisfied* **do**
  Shuffle the examples in $\mathcal{L}$
  foreach $(\mathbf{x}, \mathbf{y}) \in \mathcal{L}$ **do**
    $\hat{\mathbf{y}} = \arg\max_{\mathbf{y} \in \mathcal{C}} f(\mathbf{x}, \mathbf{y})$
    Perform graph closure on $\hat{\mathbf{y}}$
    if $\hat{\mathbf{y}} \neq \mathbf{y}$ **then**
      $\mathbf{w}_r = \mathbf{w}_r + \lambda(\sum_{i:\mathbf{y}_i = r} \phi_i - \sum_{i:\hat{\mathbf{y}}_i = r} \phi_i), \forall r \in \mathcal{Y}$
return $\{\mathbf{w}_r\}_{r \in \mathcal{Y}}$

Semi-supervised Structured Learning

- constraint-driven learning (CoDL) algorithm

Algorithm 2: Constraint-driven learning algorithm

Input: Labeled set $\mathcal{L}$, unlabeled set $\mathcal{U}$, weighting coefficient γ

Perform closure on each graph in $\mathcal{L}$

Initialize $\mathbf{w}_r = \text{Learn}(\mathcal{L})_r, \forall\, r \in \mathcal{Y}$

while *convergence criteria not satisfied* **do**

    $\mathcal{T} = \emptyset$

    foreach $\mathbf{x} \in \mathcal{U}$ **do**

        $\hat{\mathbf{y}} = \arg\max_{\mathbf{y} \in \mathcal{C}} f(\mathbf{x}, \mathbf{y})$

        Perform graph closure on $\hat{\mathbf{y}}$

        $\mathcal{T} = \mathcal{T} \cup \{(\mathbf{x}, \hat{\mathbf{y}})\}$

    $\mathbf{w}_r = \gamma \mathbf{w}_r + (1 - \gamma)\text{Learn}(\mathcal{T})_r, \forall\, r \in \mathcal{Y}$

return $\{\mathbf{w}_r\}_{r \in \mathcal{Y}}$

Missing Annotations

Table 1: Categories of E-E TLINKs in the TE3 Platinum dataset. Among all pairs of events, 98.2% of them are left unspecified by the annotators. Graph closure can automatically add 8.7%, but most of the event pairs are still unknown.

Type		#TLINK	%
Annotated		582	1.8
Missing	Inferred	2840	8.7
	Unknown	29240	89.5
Total		32662	100

- The problem of identifying these unknown relations in training and test is a major issue that dramatically hurts existing methods

- a lot of the unknown pairs are not really vague
- the scarcity of non-vague TLINKs makes it hard to learn a good `vague` classifier
- `vague` is fundamentally different from the other relation type
  - if a `before` can be established given a sentence, then it always holds as `before` regardless of other events around it
- without the `vague` classifier, the predicted temporal graph is densely connected
  - global transitivity constraints can be more effective

Vague Relation

- multiple relations can exist at this pair of events
- two annotators disagree on the relation
- human annotation
  - the annotators try to assign all possible relations to a TLINK
  - change the relation to `vague` if more than one can be assigned
  - different from many existing method
- `post-filtering` method
  - relative entropy
  - $\{r_m\}_{m=1}^{M}$ be the set of relations that i-th pair can take
  - $\delta_i = \sum_{m=1}^{M} f_{r_m}(\phi_i) \log(M f_{r_m}(\phi_i))$
  - change it to `vague` if $\delta_i \leq \tau$

Experiments

- TempEval3(TE3) workshop
- TimeBank(TB), AQUAINT(AQ),Silver(TE3-SV), Platinum(TE3-PT)
- TB, AQ for training, TE3-PT for testing
- TE3-SV is a much large, machine-annotated
- augmentations on TB
  - VerbClause (VC)
  - TimebankDense (TD)

Table 2: Facts about the datasets used in this paper. The TD dataset is split into train, dev, and test in the same way as in Chambers et al. (2014). Note that the column of TLINKs only counts the non-vague TLINKs, from which we can see that the TD dataset has a much higher ratio of #TLINKs to #Events. The TLINK annotations in TE3-SV is not used in this paper and its number is thus not shown.

Dataset	Doc	Event	TLINK	Note
TB+AQ	256	12K	12K	Training
VC	132	1.6K	0.9K	Training
TD	36	1.6K	5.7K	Training
TD-Train	22	1K	3.8K	Training
TD-Dev	5	0.2K	0.6K	Dev
TD-Test	9	0.4K	1.3K	Eval
TE3-PT	20	0.7K	0.9K	Eval
TE3-SV	2.5K	81K	-	Unlabeled

TE3 Task C - Relation Only

Table 3: Temporal awareness scores on TE3-PT given gold event pairs. Systems that are significantly better (per McNemar's test with $p < 0.0005$) than the previous rows are underlined. The last column shows the relative improvement in F1 score over AP-1, which identifies the source of improvement: 5.2% from additional training data, 9.3% (14.5%-5.2%) from constraints, and 10.4% from structured learning.

Method	P	R	F1	%
UTTime	55.6	57.4	56.5	+5.0
AP-1	56.3	51.5	53.8	0
AP-2	58.0	55.3	56.6	+5.2
AP+ILP	62.2	61.1	61.6	+14.5
SP+ILP	**69.1**	**65.5**	**67.2**	**+24.9**

- AP: regularized average perceptron
- SP: structured perceptron

Table 4: Temporal awareness scores given gold events but with no gold pairs, which show that the proposed S+I methods outperformed state-of-the-art systems in various settings. The fourth column indicates the annotation sources used, with additional unlabeled dataset in the parentheses. The "Filters" column shows if the pre-filtering method (Sec. 3.1) or the proposed post-filtering method (Sec. 4) were used. The last column is the relative improvement in F_1 score compared to baseline systems on line 1, 7, and 11, respectively. Systems that are significantly better than the "*"-ed systems are underlined (per McNemar's test with $p < 0.0005$).

No.	System	Method	Anno. (Unlabeled)	Testset	Filters	P	R	F1	%
1	ClearTK	Local	TB, AQ, VC, TD	TE3-PT	pre	**37.2**	33.1	35.1	0
2	AP*	Local	TB, AQ, VC, TD	TE3-PT	pre	35.3	37.1	36.1	+2.8
3	AP+ILP	L+I	TB, AQ, VC, TD	TE3-PT	pre	35.7	35.0	35.3	+0.6
4	SP+ILP	S+I	TB, AQ, VC, TD	TE3-PT	pre	32.4	45.2	37.7	+7.4
5	SP+ILP	S+I	TB, AQ, VC, TD	TE3-PT	pre+post	33.1	**49.2**	39.6	+12.8
6	CoDL+ILP	S+I	TB, AQ, VC, TD (TE3-SV)	TE3-PT	pre+post	35.5	46.5	**40.3**	**+14.8**
7	ClearTK*	Local	TB, VC	TE3-PT	pre	**35.9**	38.2	37.0	0
8	SP+ILP	S+I	TB, VC	TE3-PT	pre+post	30.7	**47.1**	37.2	+0.5
9	CoDL+ILP	S+I	TB, VC (TE3-SV)	TE3-PT	pre+post	33.9	45.9	**39.0**	**+5.4**
10	ClearTK	Local	TD-Train	TD-Test	pre	46.04	20.90	28.74	-
11	CAEVO*	L+I	TD-Train	TD-Test	pre	**54.17**	39.49	45.68	0
12	SP+ILP	S+I	TD-Train	TD-Test	pre+post	45.34	48.68	46.95	+3.0
13	CoDL+ILP	S+I	TD-Train (TE3-SV)	TD-Test	pre+post	45.57	**51.89**	**48.53**	**+6.3**

Conclusion

- structured learning approach
- capturing the global nature
- a new perspective towards `vague` relations
- making use of the unlabeled data
- improved performance on both TE3-PT and TD-Test

Thanks Q&A