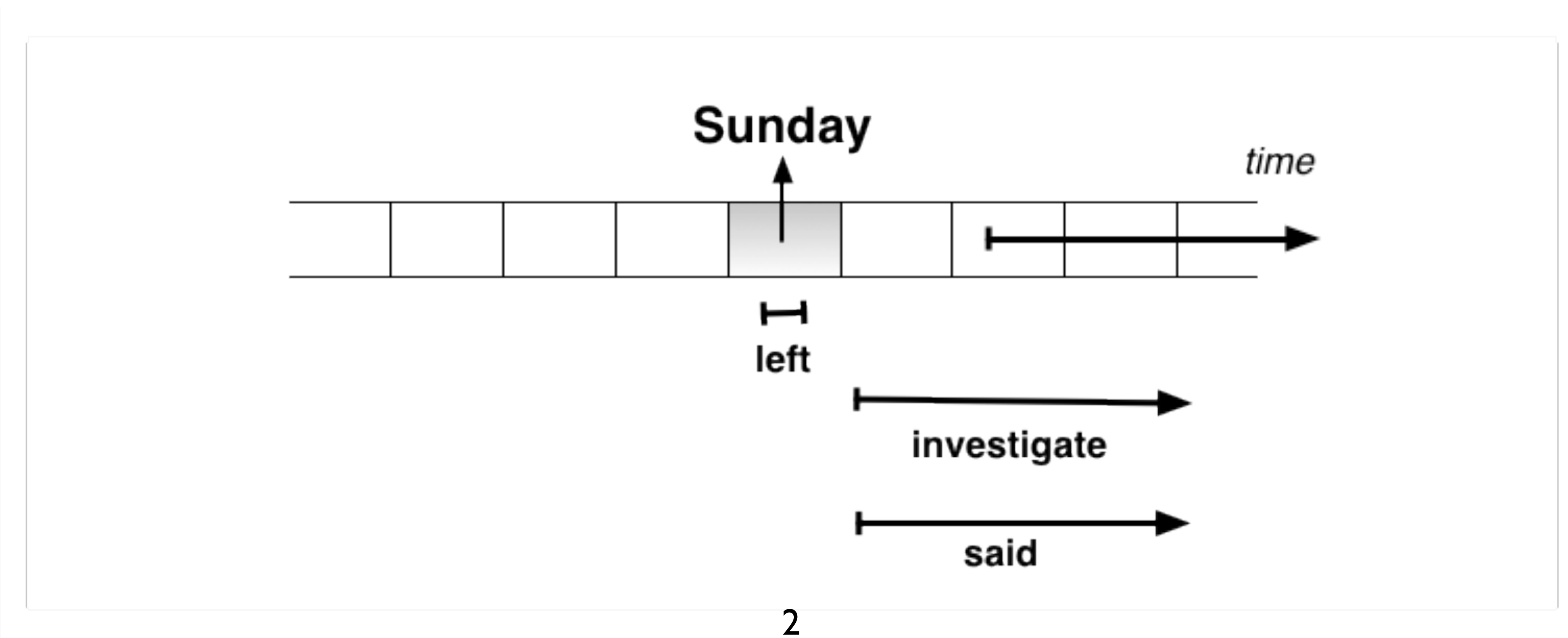
Improved Temporal Relation Classification

using Dependency Parses and Selective Crowdsourced Annotations

Jun-Ping Ng and Min-Yen Kan

Temporal Relations?

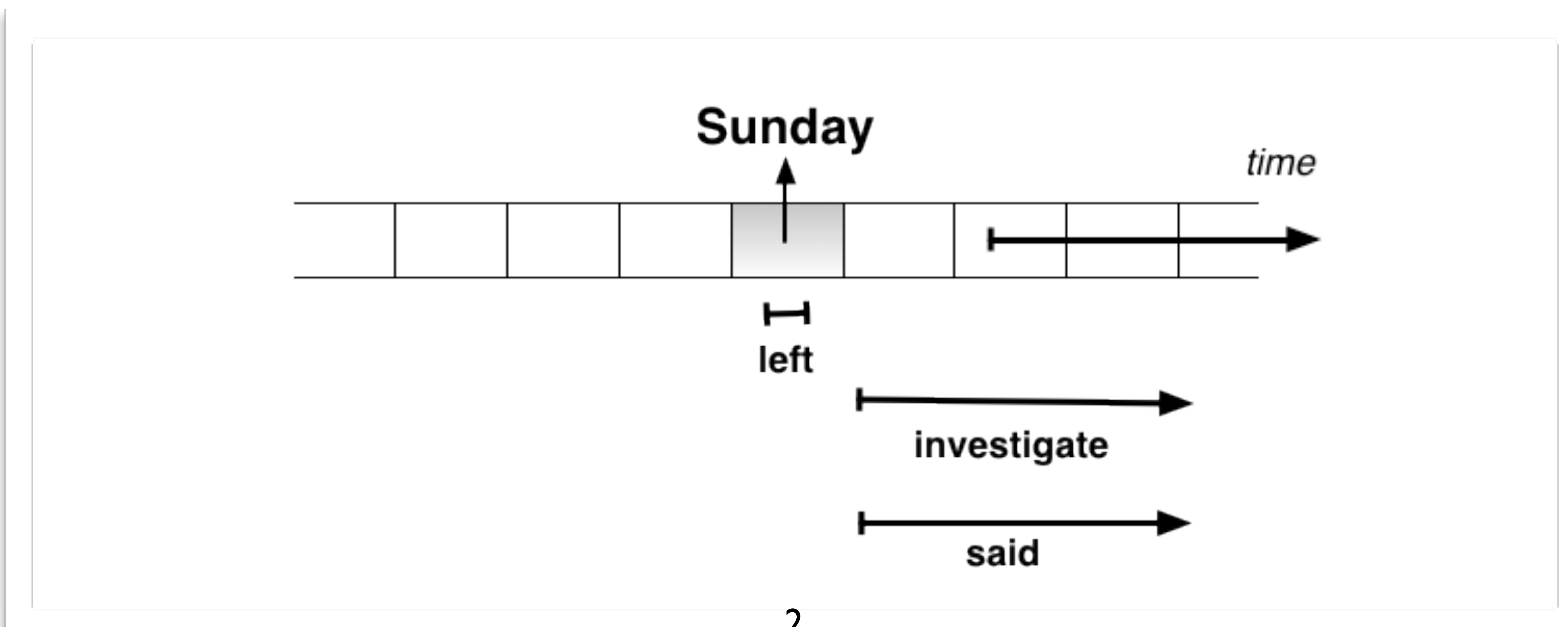
Two top aides to Netanyahu, political advisor Uzi Arad and Cabinet Secretary Danny Naveh, **left** for Europe on Sunday, apparently to **investigate** the Syrian issue, the newspaper said.



Temporal Relations?

OVERLAP

Two top aides to Netanyahu, political advisor Uzi Arad and Cabinet Secretary Danny Naveh, **left** for Europe on Sunday, apparently to **investigate** the Syrian issue, the newspaper said.

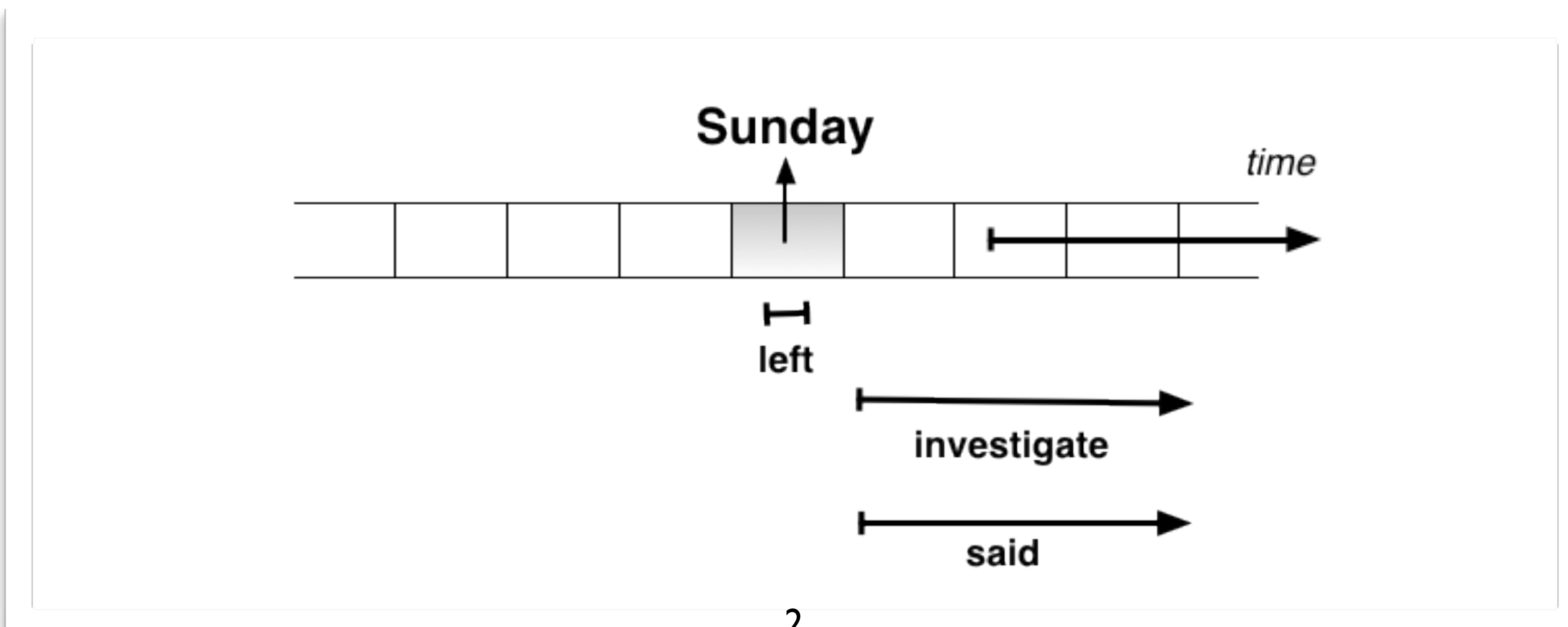


Temporal Relations?

OVERLAP

Two top aides to Netanyahu, political advisor Uzi Arad and Cabinet Secretary Danny Naveh, **left** for Europe on Sunday, apparently to **investigate** the Syrian issue, the newspaper said.

AFTER



Goal

- Be able to classify event-temporal (E-T) relations within a sentence

Outline

- Brief look at state-of-the-art
- Proposed Approach
  - Reducing size of feature space
  - Smart acquisition of data via crowdsourcing
- Error Analysis

State-of-the-art

- Shared tasks TempEval-1 and TempEval-2 held in conjunction with SemEval in 2007 and 2010.
- State-of-the-art systems in TempEval-2 achieve around 65% accuracy
- Work with dataset from TempEval-2 to facilitate benchmarking and comparisons

Data Sparsity

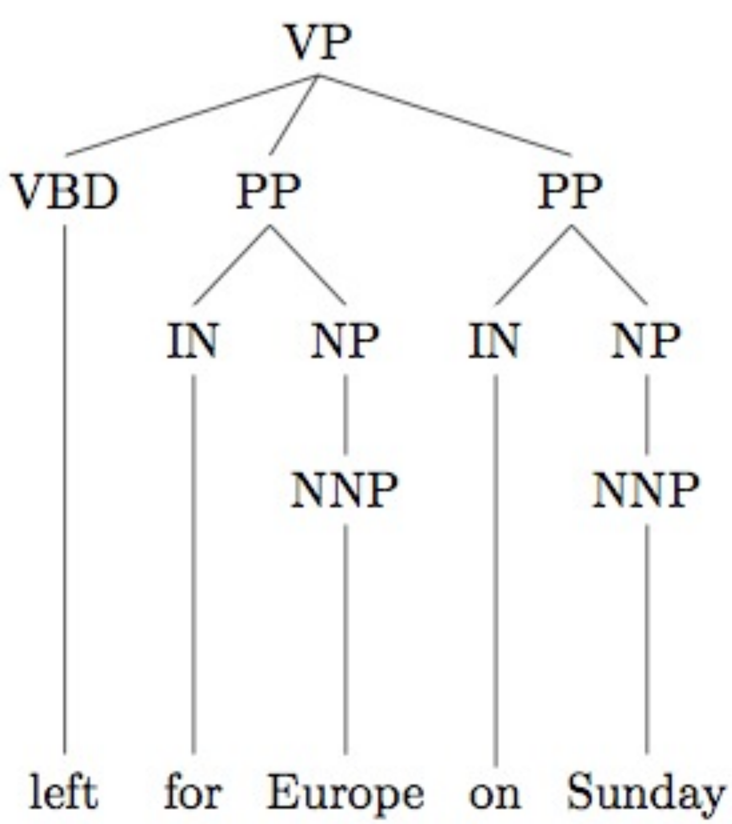
- Features typically employed include
  - lexical cues
  - context
  - sentence structure
- Training set consists of around 959 instances

Proposal

- Reduce dimensionality of feature space
- Increase amount of annotated data available

A Kernel Hypothesis

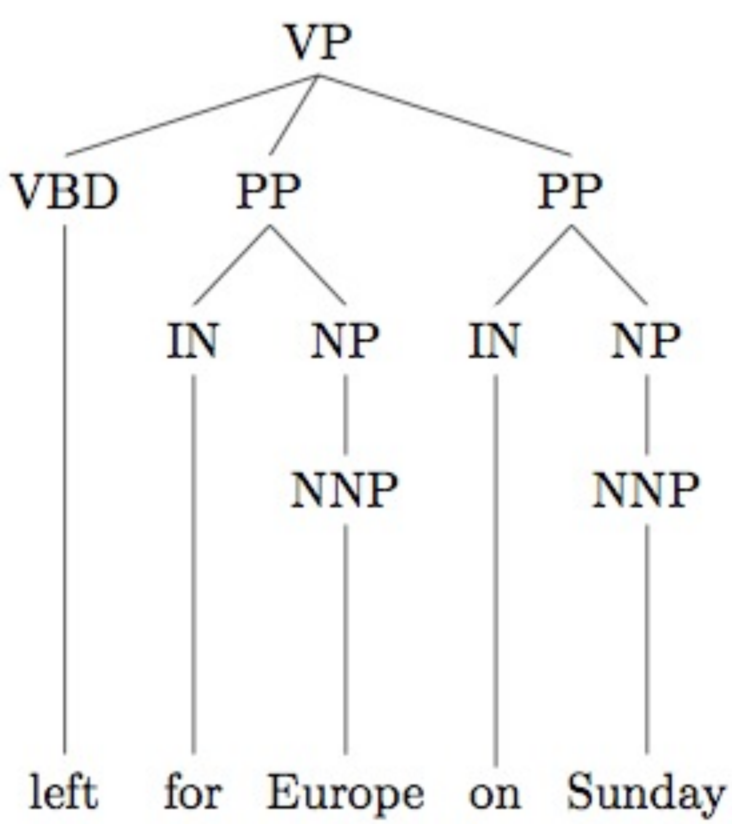
... **left** for Europe on <u>Sunday</u> ...



A Kernel Hypothesis

OVERLAP

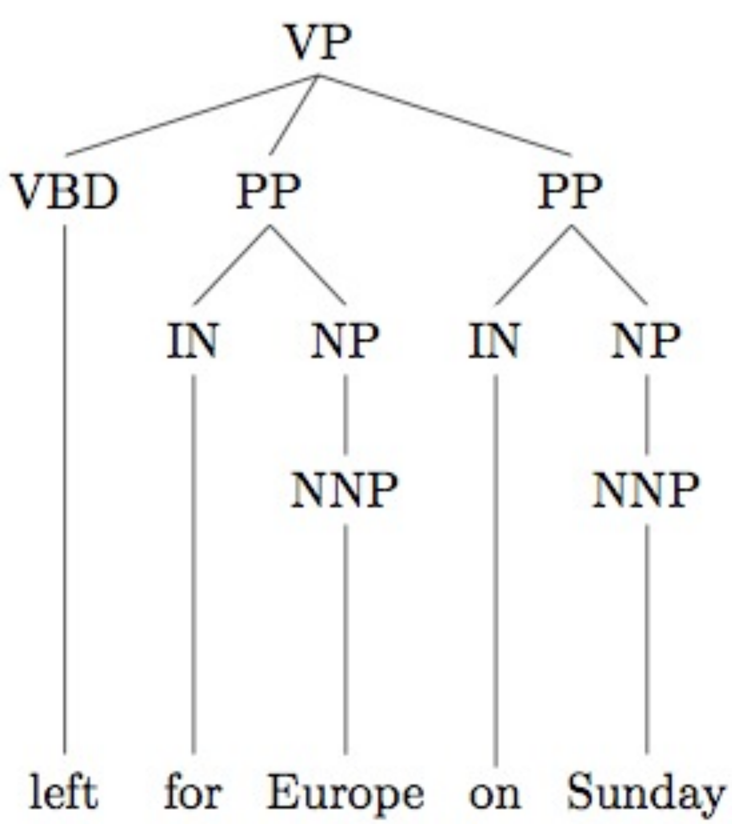
... **left** for Europe on Sunday ...



A Kernel Hypothesis

OVERLAP

... **left** for Europe on Sunday ...



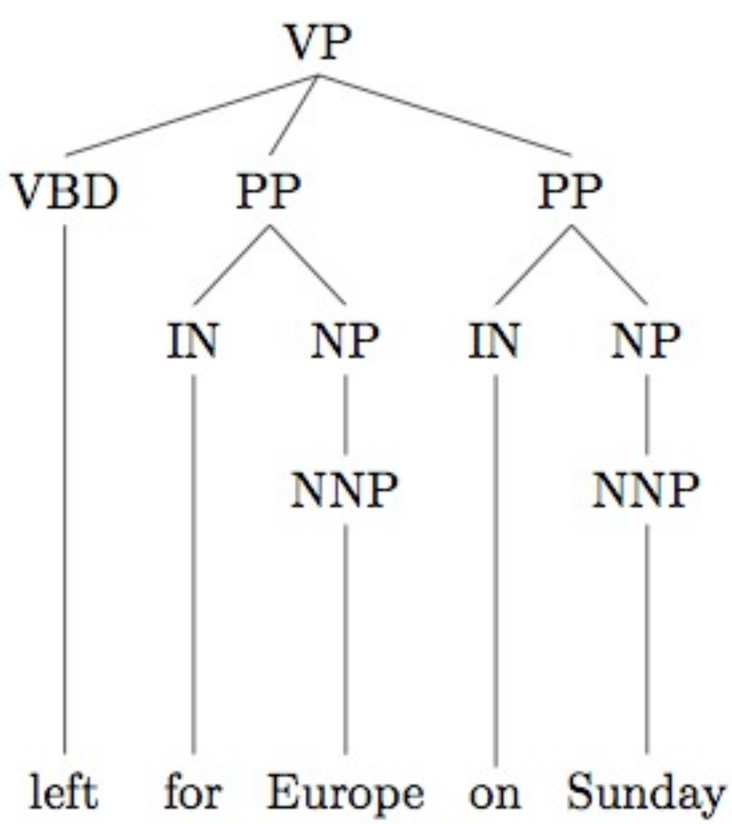
...**went** to America on Monday..

...**partied** at home on Wednesday..

A Kernel Hypothesis

OVERLAP

... **left** for Europe on Sunday ...



OVERLAP

...**went** to America on Monday..

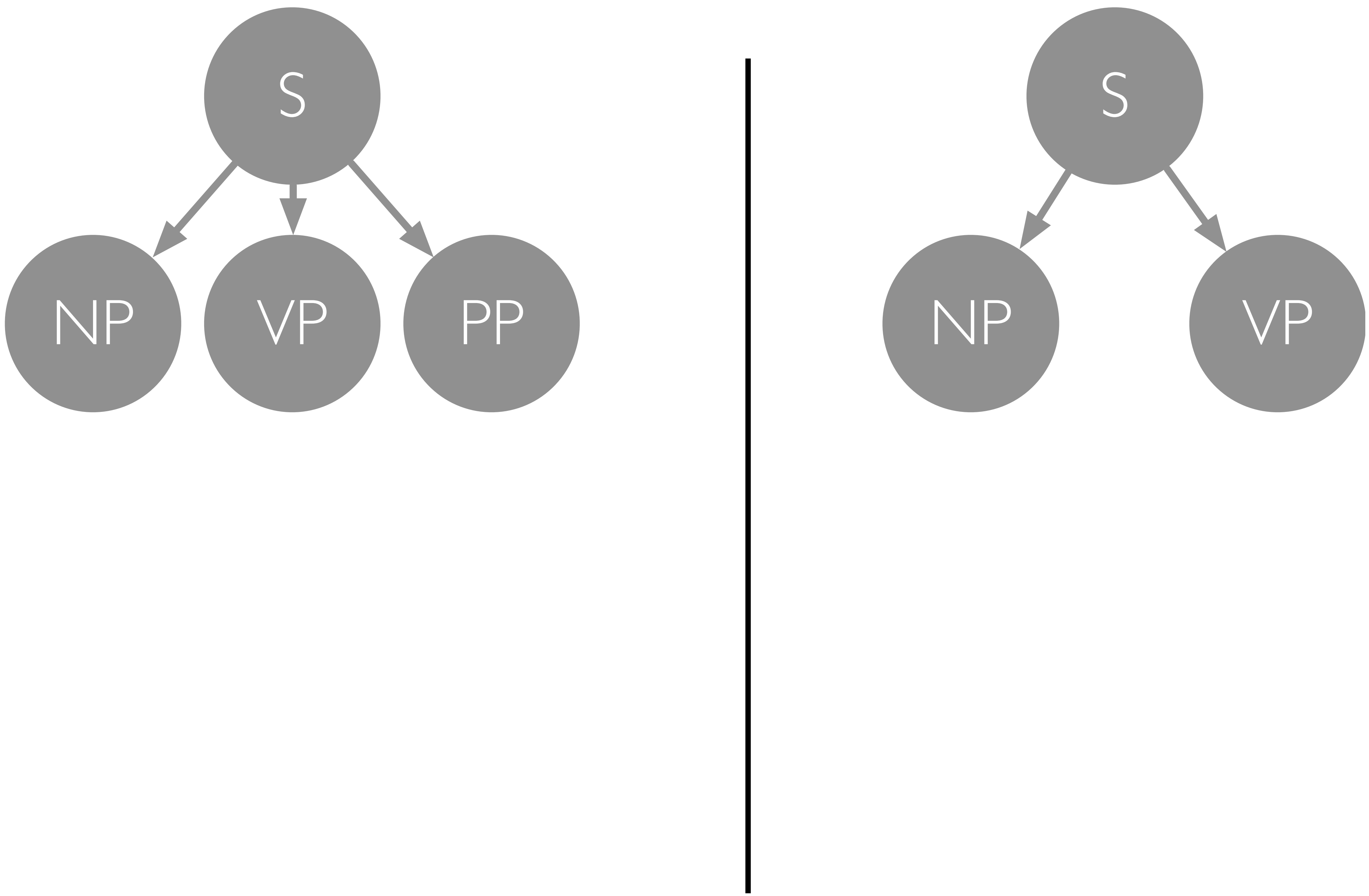
...**partied** at home on Wednesday..

OVERLAP

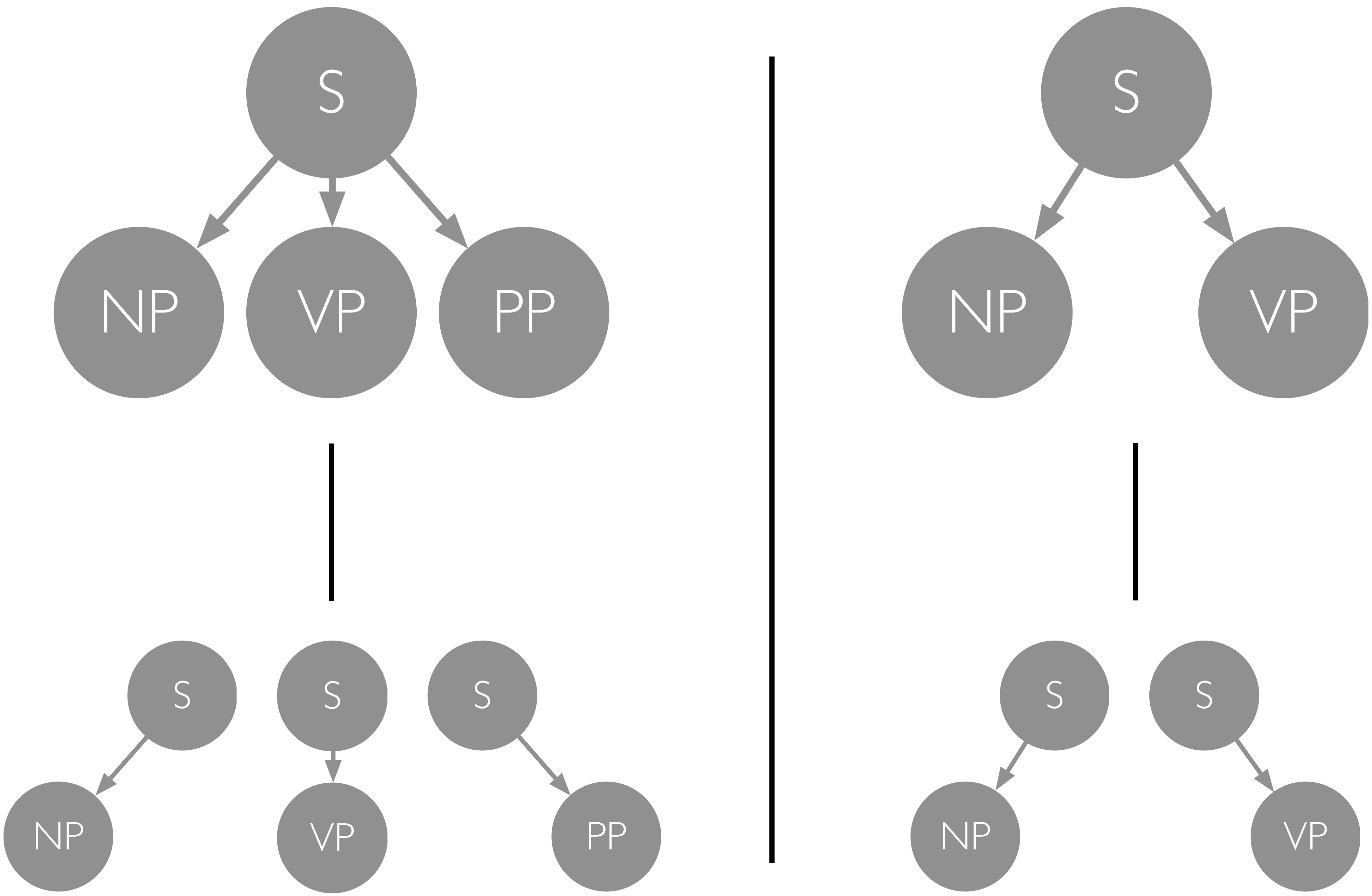
Convolution Kernels

- Allows us to capture structure similarities easily
- Tree structure used as feature for support vector machines (SVM)
- No need to flatten structure representation into a set of real number features

Structure Similarity

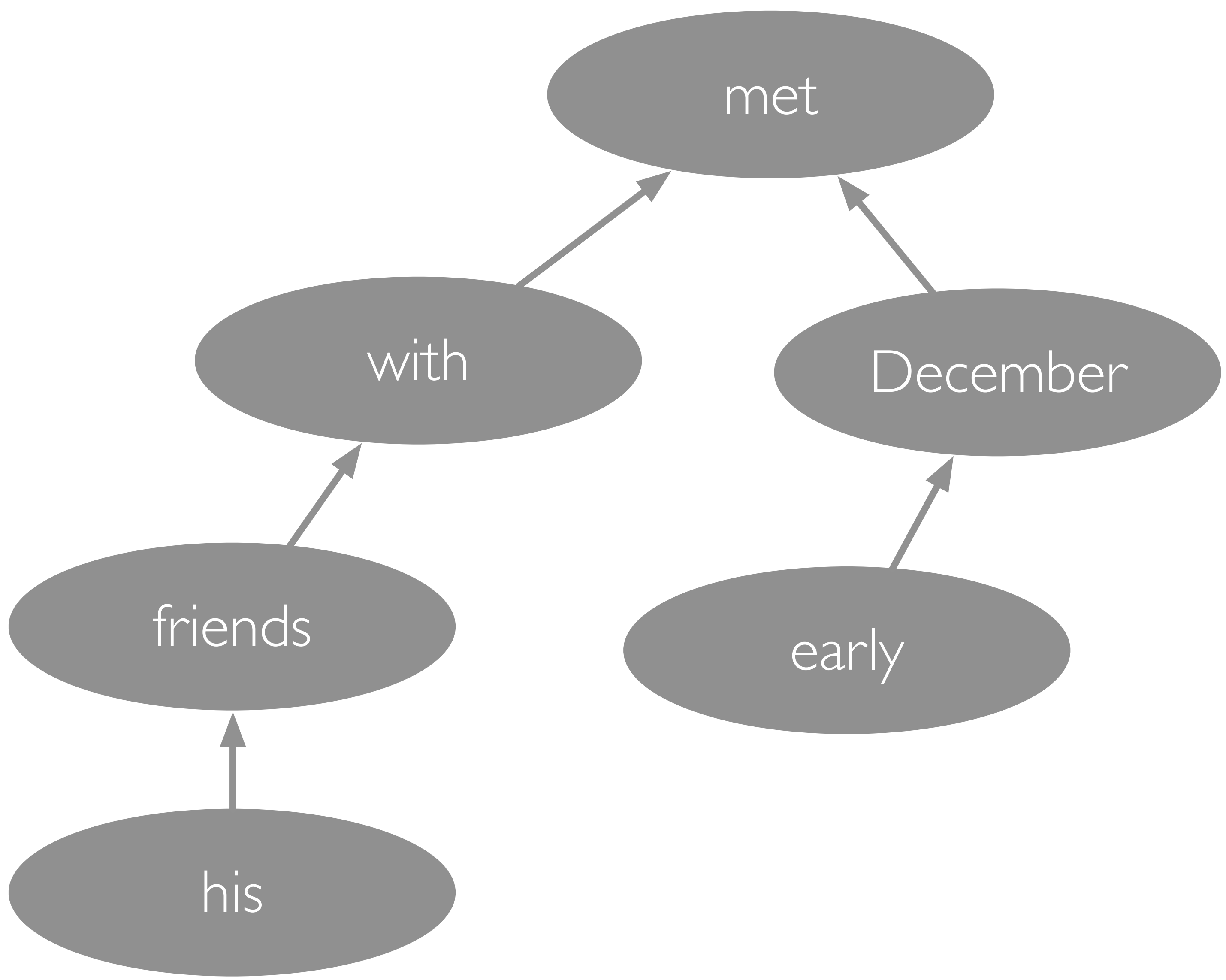


Structure Similarity



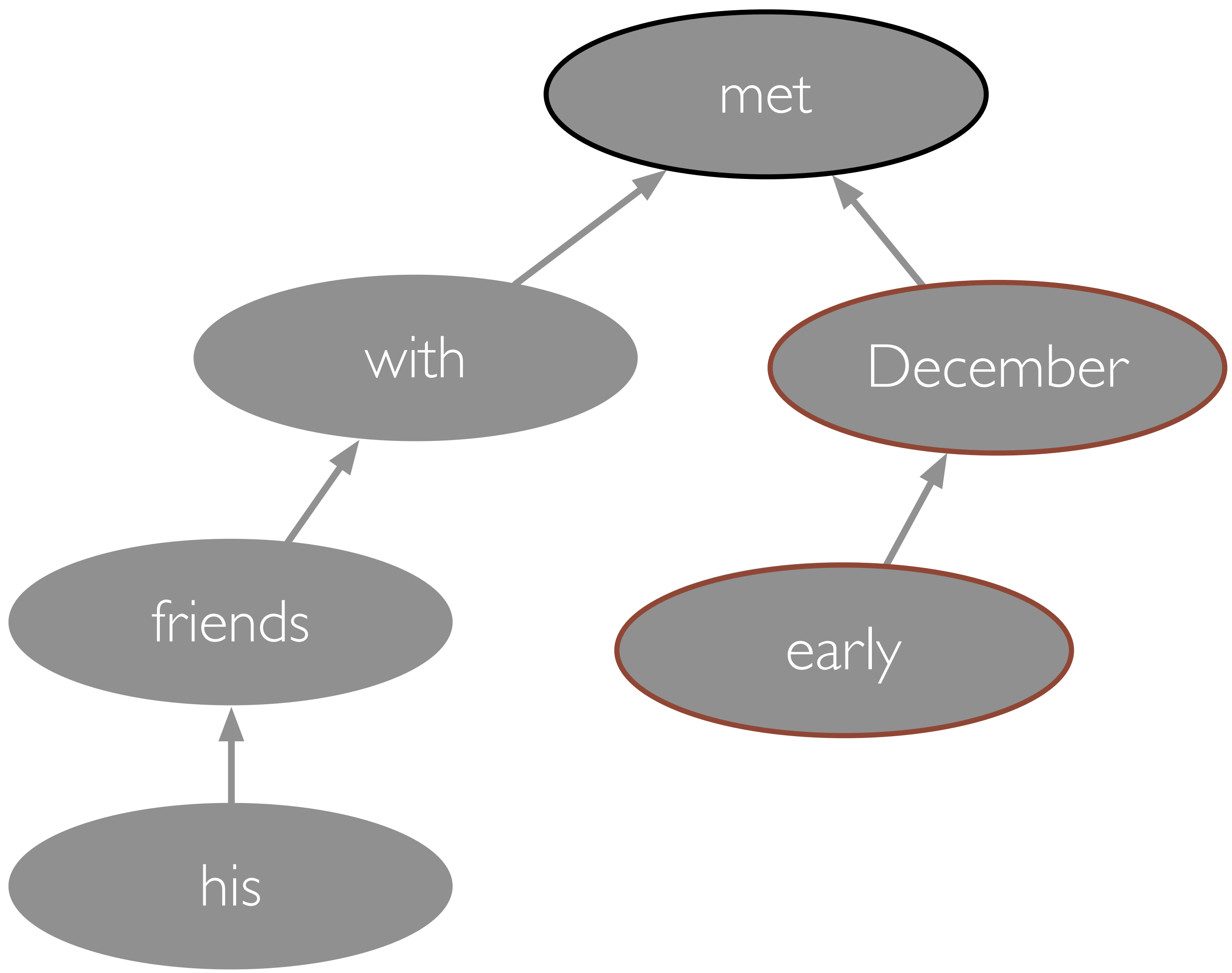
Features

... met with his friends early December ...



Features

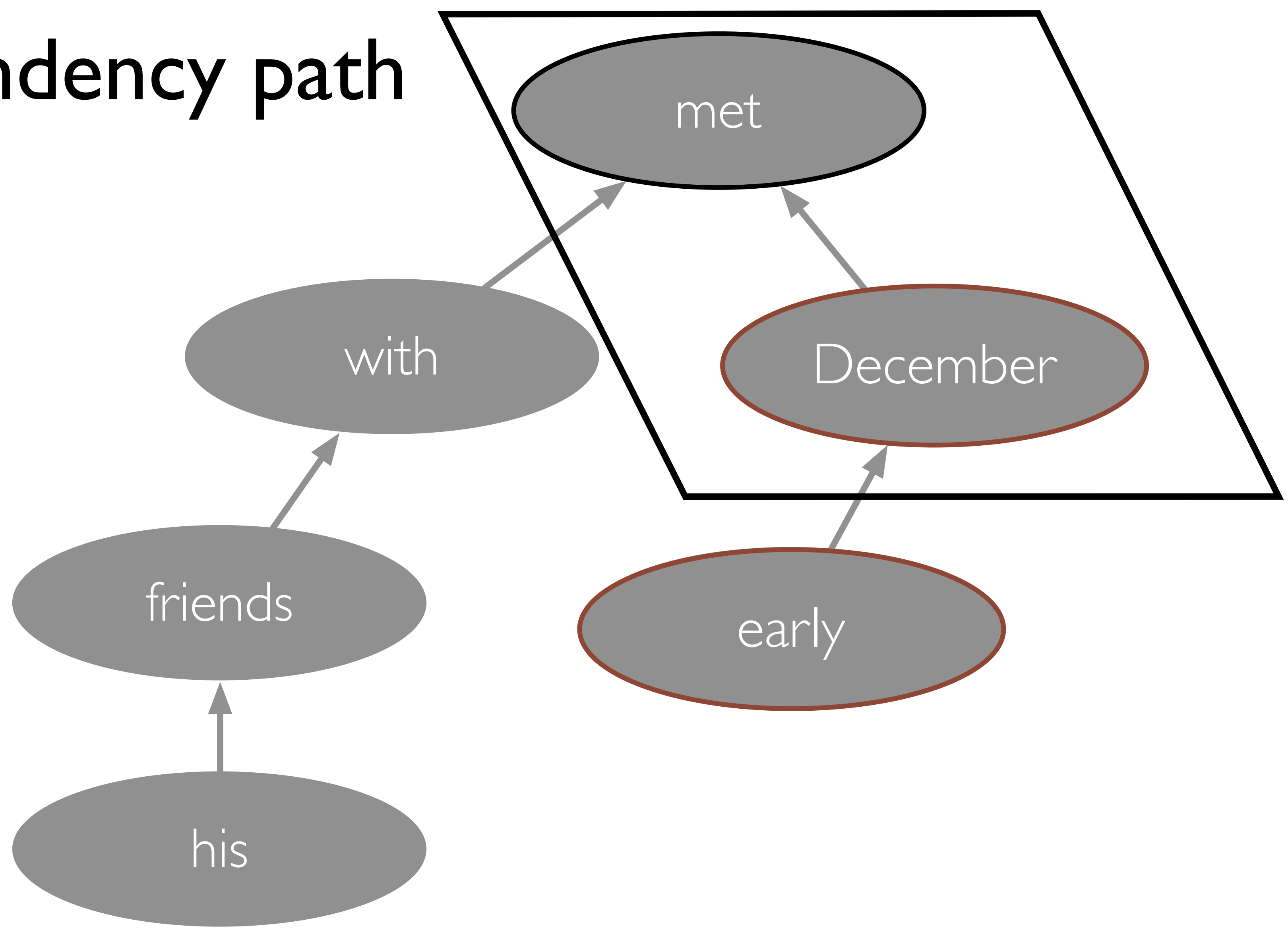
... met with his friends early December ...



Features

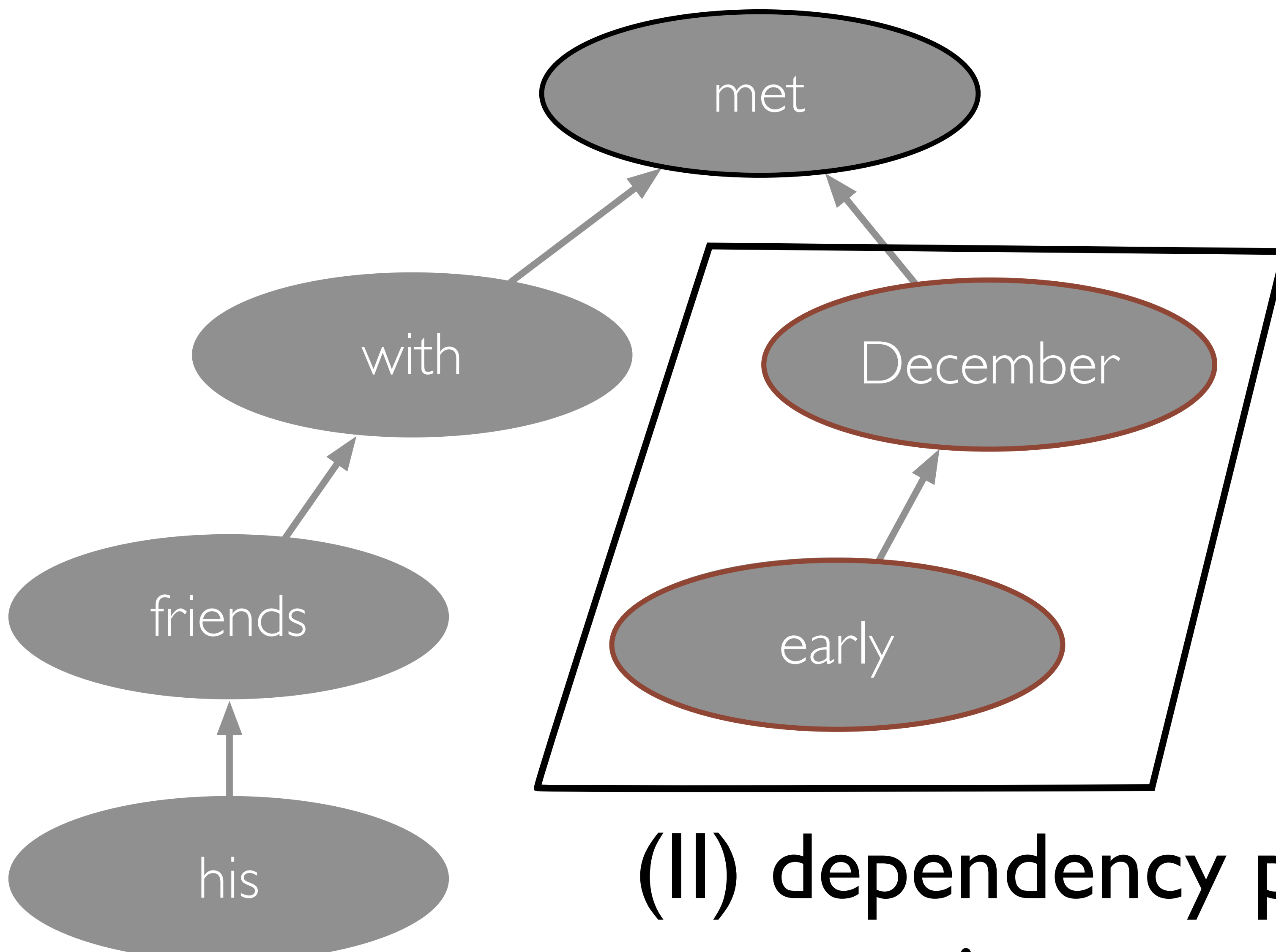
... met with his friends early December ...

(1) dependency path



Features

... met with his friends early December ...



(II) dependency parse of time expression

Comparisons

System	Accuracy
ConvoDep	67.4%
TRIOS	65.0%
JU_CSE	63.0%
NCSU_indi	63.0%
NCSU_joint	63.0%
TRIPS	63.0%
USFD2	63.0%

- Trained on TempEval-2 training set, Tested on TempEval-2 testing set

Comparisons

System	Accuracy
ConvoDep	67.4%
TRIOS	65.0%
JU_CSE	63.0%
NCSU_indi	63.0%
NCSU_joint	63.0%
TRIPS	63.0%
USFD2	63.0%

Precision	0.828
Recall	0.512
F1	0.523

✦ Trained on TempEval-2 training set, Tested on TempEval-2 testing set

Getting More Data

- Crowdsourcing is a cheap, efficient avenue for large scale data annotation
- But temporal annotations are not trivial
- We want to investigate
  - the quality of crowdsourced temporal annotations
  - effective ways to gather the annotations

Task Setup

- Crowdsourcing via Crowdflower
  - Data validation to improve data quality
- Raw data collected from news articles
  - Event and time expressions extracted during pre-processing

Is It Useful?

- Collected initial dataset of 1000 instances
- Trained SVM classifier with convolution kernels

Is It Useful?

System	Accuracy	F1	Precision	Recall
ConvoDep	67.4%	0.523	0.828	0.512
CF-1000	65.2%	0.525	0.578	0.535
CF-1000 + TE	71.7%	0.615	0.726	0.598

✦ Tested on TempEval-2 testing set

A Smarter Way

- Are we able to collect less data but still remain effective?
- Insight - Instances are not equally hard

A Smarter Way

- Are we able to collect less data but still remain effective?
- Insight - Instances are not equally hard

Two top aides to Netanyahu, political advisor Uzi Arad and Cabinet Secretary Danny Naveh, **left** for Europe on <u>Sunday</u>, apparently to investigate the Syrian issue, the newspaper **said**.

A Smarter Way

- Are we able to collect less data but still remain effective?
- Insight - Instances are not equally hard

Two top aides to Netanyahu, political advisor Uzi Arad and Cabinet Secretary Danny Naveh, **left** for Europe on <u>Sunday</u>, apparently to investigate the Syrian issue, the newspaper **said**.

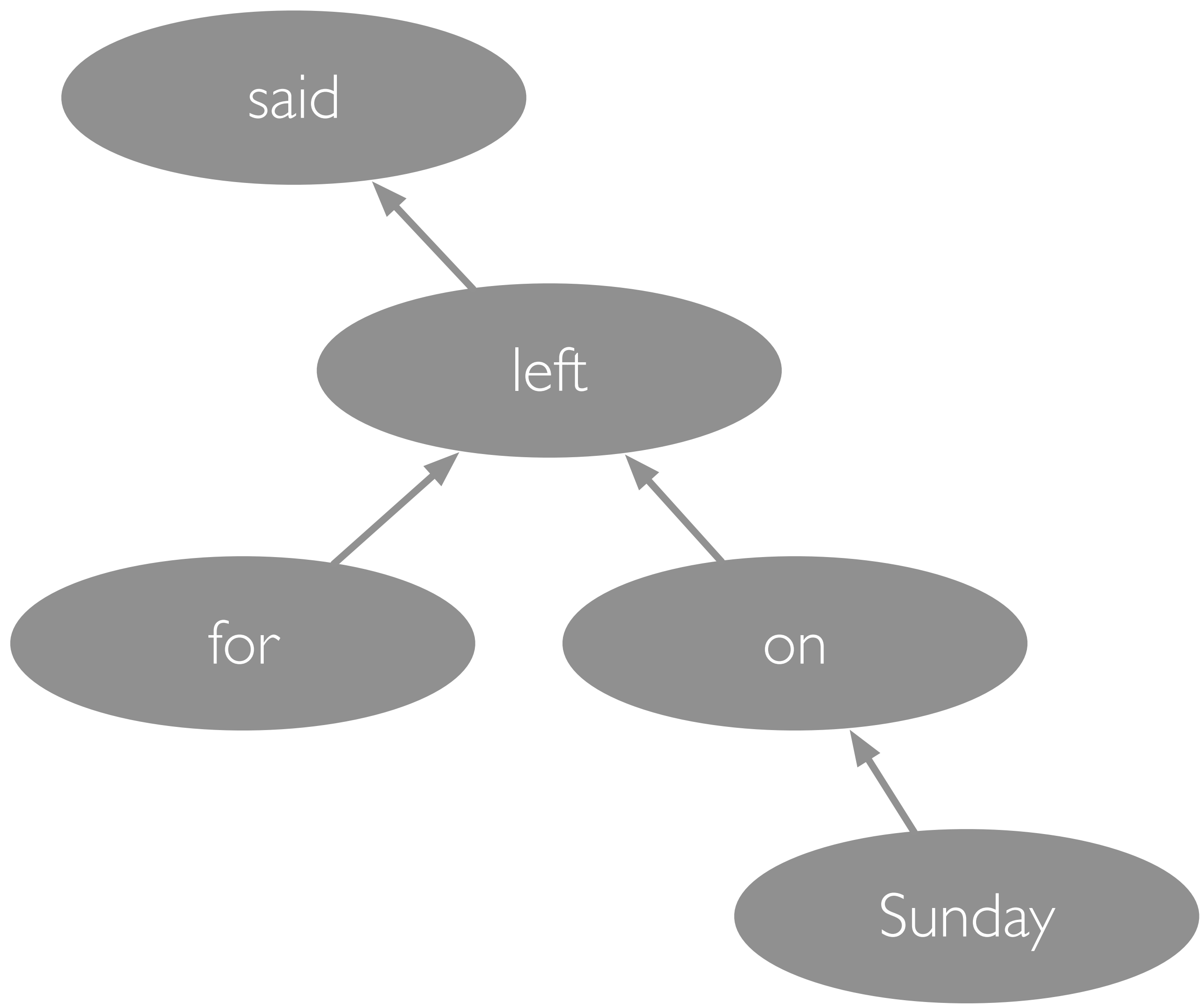
A Smarter Way

- Are we able to collect less data but still remain effective?
- Insight - Instances are not equally hard

Two top aides to Netanyahu, political advisor Uzi Arad and Cabinet Secretary Danny Naveh, **left** for Europe on Sunday, apparently to investigate the Syrian issue, the newspaper **said**.

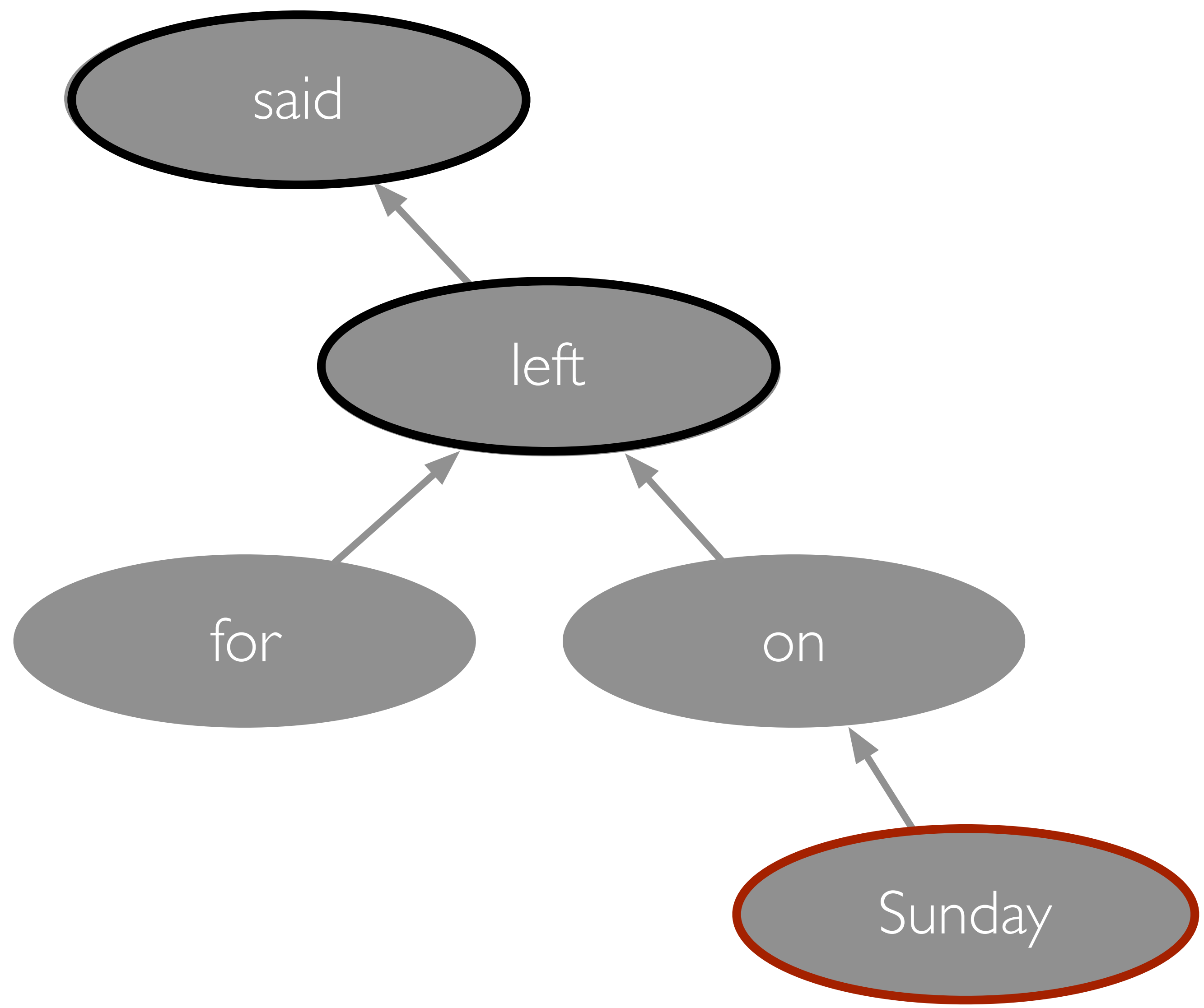
Hard Instances

- Order event expressions in increasing order from time expression in dependency parse



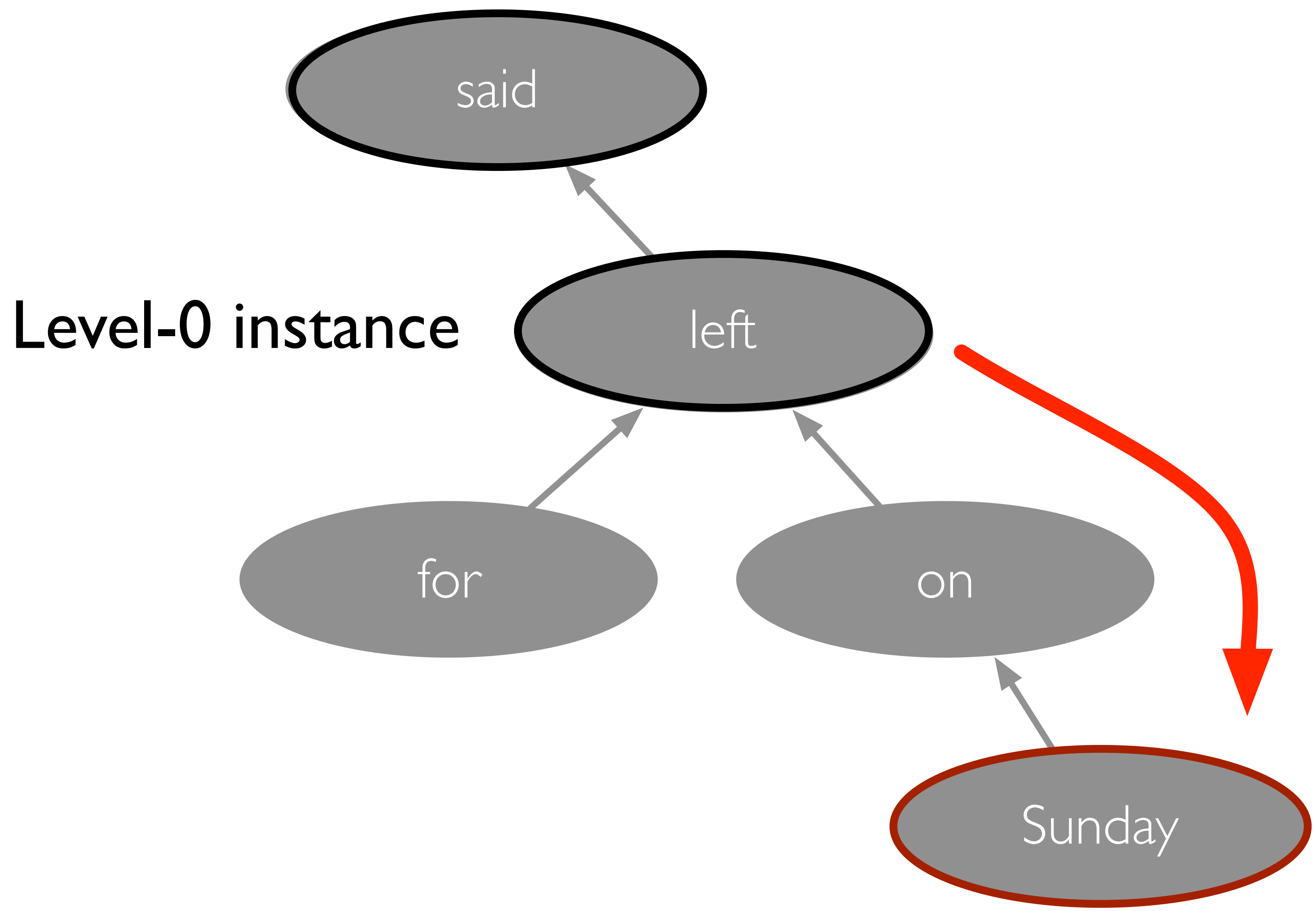
Hard Instances

- Order event expressions in increasing order from time expression in dependency parse



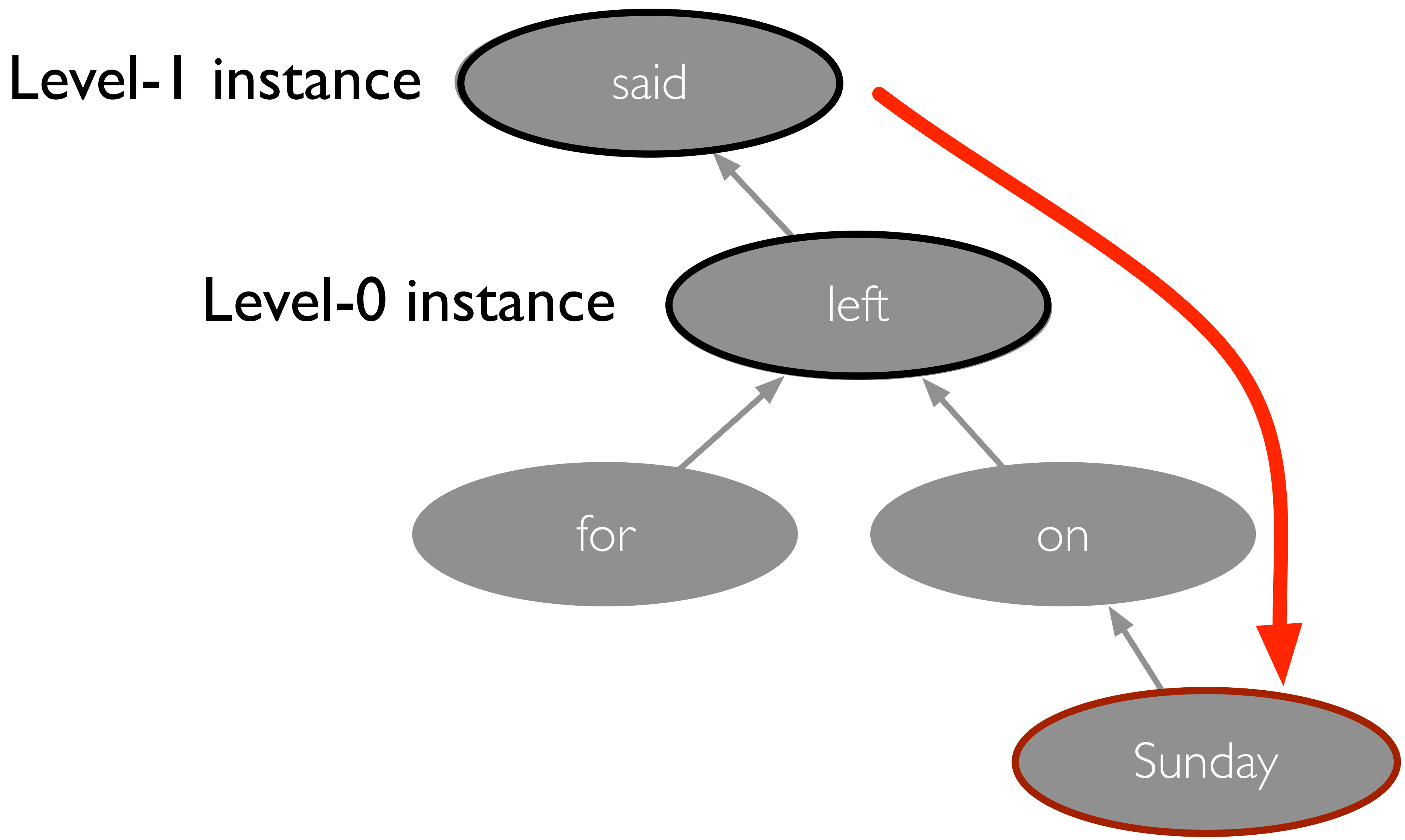
Hard Instances

- Order event expressions in increasing order from time expression in dependency parse



Hard Instances

- Order event expressions in increasing order from time expression in dependency parse



Easy Level-0 Instances

- Level-0 instances are much easier to get correct

Accuracy (%)				
Level-0 (59)	Level-1 (47)	Level-2 (21)	Level-3 (10)	Level-4 (1)
84.5	66.0	42.9	30.0	100.0

✦ Tested on TempEval-2 testing set

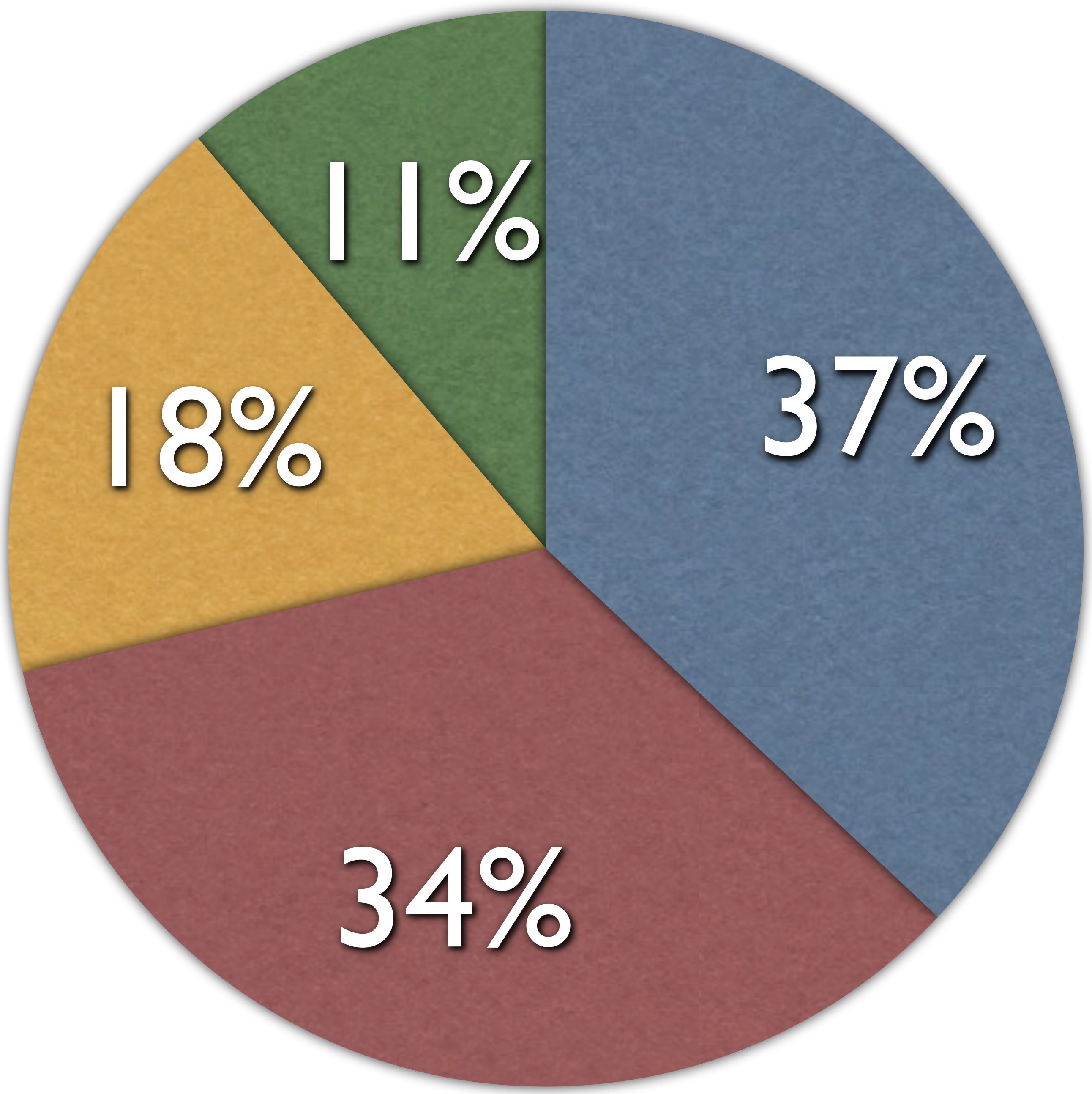
Selective Annotation

- Dropping Level-0 instances does not lead to drop in performance

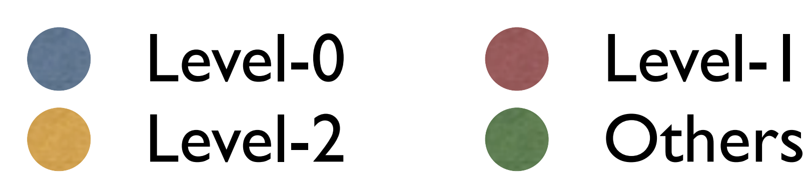
System	Accuracy	F1	Precision	Recall
ConvoDep	67.4%	0.523	0.828	0.512
CF-NoLevel0	73.2%	0.639	0.659	0.643
CF-Full	73.2%	0.641	0.660	0.647

✦ Tested on TempEval-2 testing set

Annotation Savings



- Level-0 instances form up to 37% of the annotated data



Analysis

- Why missing out on 37% of training instances causes no drop in performance?
- How to approach performance upper-bound?

Performance Breakdown

- Classifier does better on OVERLAP relations

System	Overlap			Before			After		
	P	R	F1	P	R	F1	P	R	F1
CF-NoLevel0	0.72	0.96	0.82	0.56	0.45	0.50	0.70	0.52	0.60
CF-Full	0.72	0.95	0.81	0.57	0.40	0.47	0.70	0.60	0.64

Label Distribution

- Level-0 instances contain less AFTER and BEFORE instances

Label	Distribution of Labels (%)		
	Level-0	Level-1	Level-2
AFTER	10.1	21.2	23.6
BEFORE	5.1	13.7	16.1

Label Distribution

- Level-0 instances contain less AFTER and BEFORE instances

Label	Distribution of Labels (%)		
	Level-0	Level-1	Level-2
AFTER	10.1	21.2	23.6
BEFORE	5.1	13.7	16.1

Confusion Matrix

- BEFORE mis-classified as AFTER

Actual Label	Predicted Label		
	OVERLAP	BEFORE	AFTER
OVERLAP	78	2	1
BEFORE	7	9	4
AFTER	13	0	14

✦ Confusion matrix for CF-NoLevel0

Confusion Matrix

- BEFORE mis-classified as AFTER

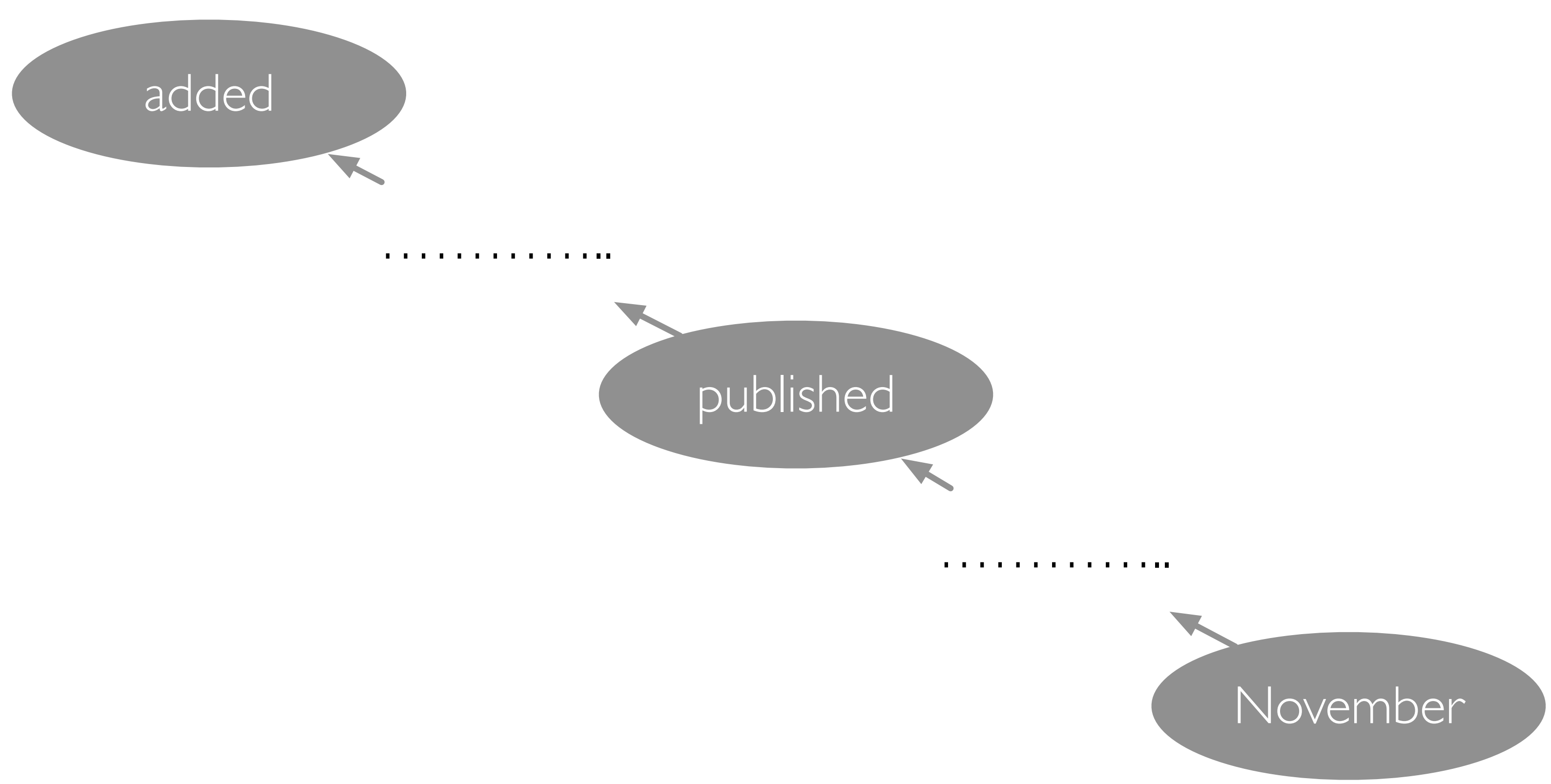
Actual Label	Predicted Label		
	OVERLAP	BEFORE	AFTER
OVERLAP	78	2	1
BEFORE	7	9	4
AFTER	13	0	14

✦ Confusion matrix for CF-NoLevel0

Copula Modifiers

BEFORE

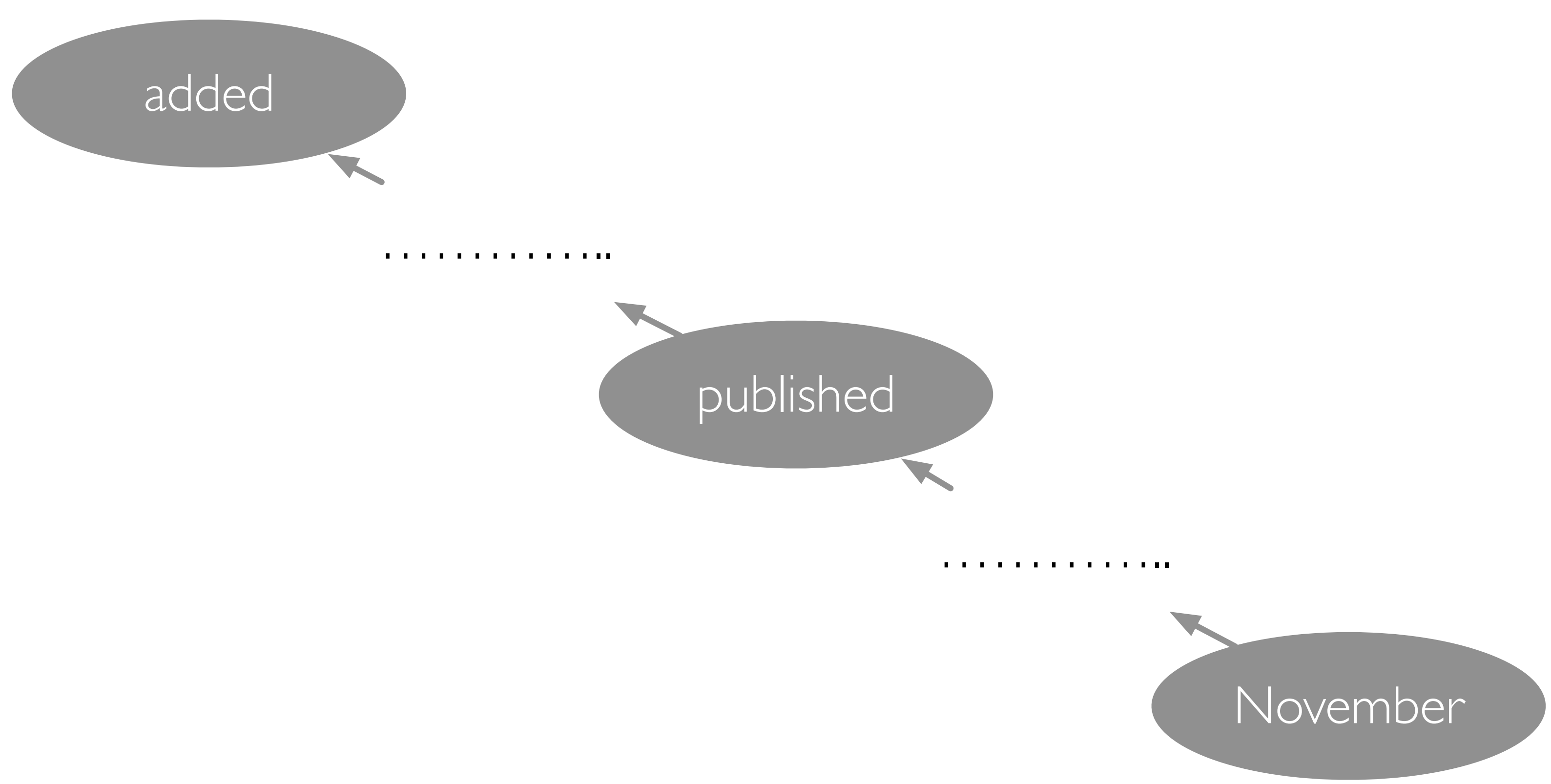
He **added** that final guidelines to be published in early <u>November</u> will determine whether the bank is in compliance.



Copula Modifiers

BEFORE

He **added** that final guidelines to be published in early <u>November</u> will determine whether the bank is in compliance.



Missing Feature

He **added** that final guidelines published in early <u>November</u> will determine whether the bank is in compliance.

Missing Feature

AFTER

He **added** that final guidelines published in early <u>November</u> will determine whether the bank is in compliance.

Round Up

- Improved event-temporal relation classification
  - By reducing input feature space
  - By increasing amount of annotated data
    - Demonstrated efficacy of crowdsourced annotations
    - Proposed an optimization to reduce the annotation effort required

Thank you!

Questions?

Dataset can be downloaded at

http://wing.comp.nus.edu.sg/~junping/etrcc/page/index.html

Breakdown of Data

Data Set	Relative size of partition (%)			
	Level-0	Level-1	Level-2	Others
TempEval-2 Training	40.9	35.2	15.1	8.8
TempEval-2 Testing	41.4	34.3	15.7	8.6
CF-Full	37.0	34.3	17.5	11.2