LANGUAGE COMPUTER

Event Detection and Coreference

TAC KBP 2015

Sean Monahan, Michael Mohler, Marc Tomlinson

Amy Book, Mary Brunson, Maxim Gorelkin, Kevin Crosby

Overview

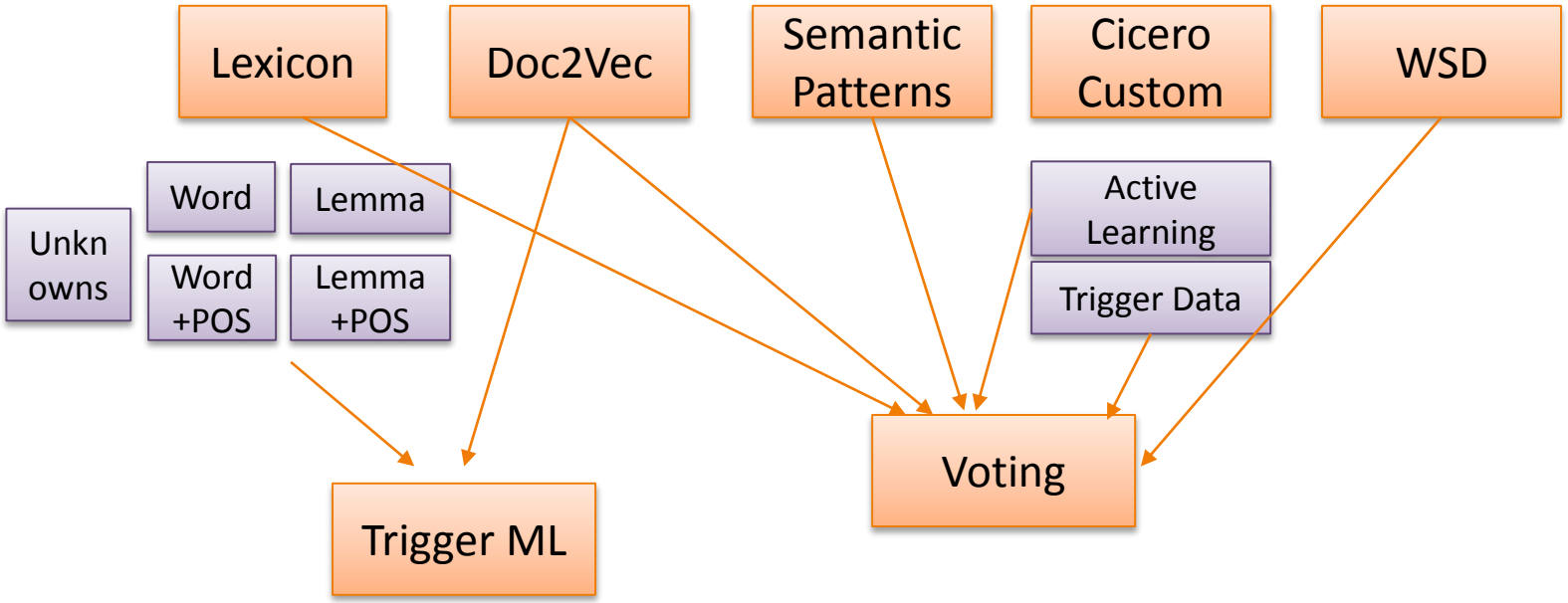
- Event Detection (Task 1)
  - What worked and what didn't
  - Lexical Knowledge
  - Annotation Ideas
- Event Hoppers (Task 2 / 3)

Event Detection – Problem Description

- Find the text which indicates the event
  - Triggers
    - "Find the smallest extent of text (usually a word or short phrase) that expresses the occurrence of an event)"
  - Nugget
    - Find the maximal extent of a textual event indicator
- Event Types
  - 38 different event types (subtypes)
  - Each with a different definition and different requirements
    - Highly varying performance per type
- Difficult Cases
  - Unclear context – "The politician attacked his rivals"
  - Unclear event – "There's murder in his blood"

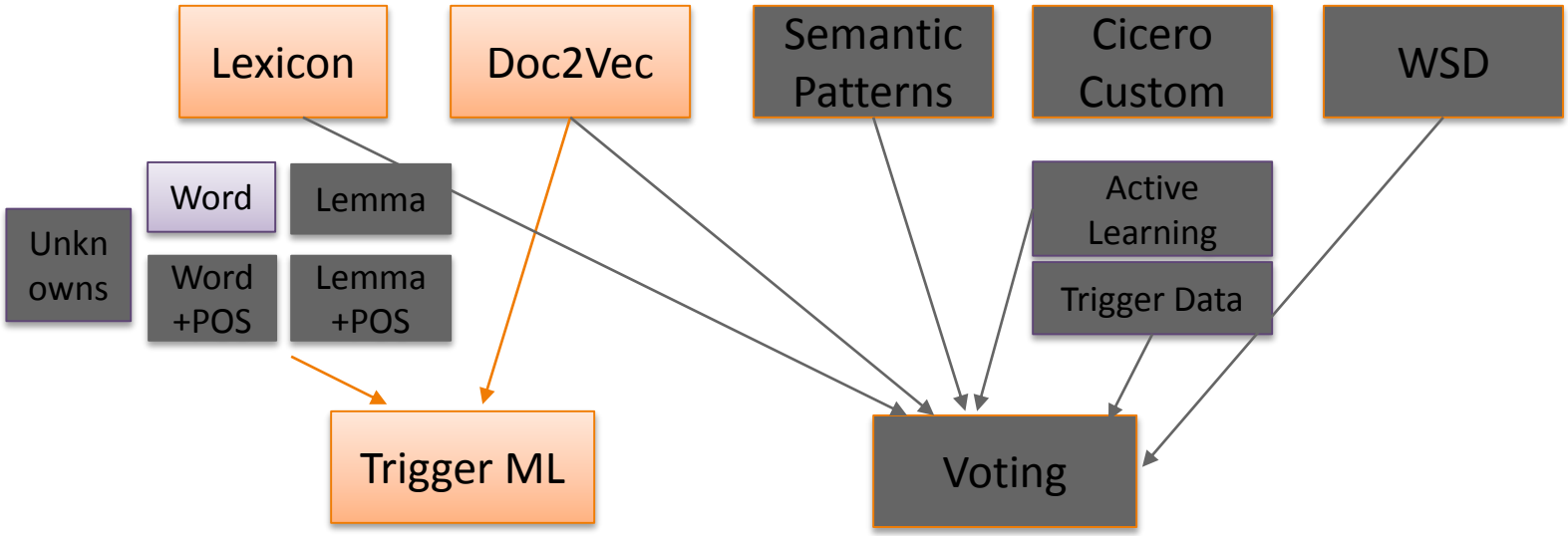
Event Detection – All Strategies

- We experimented with a lot of different strategies



Event Detection – Working Strategies

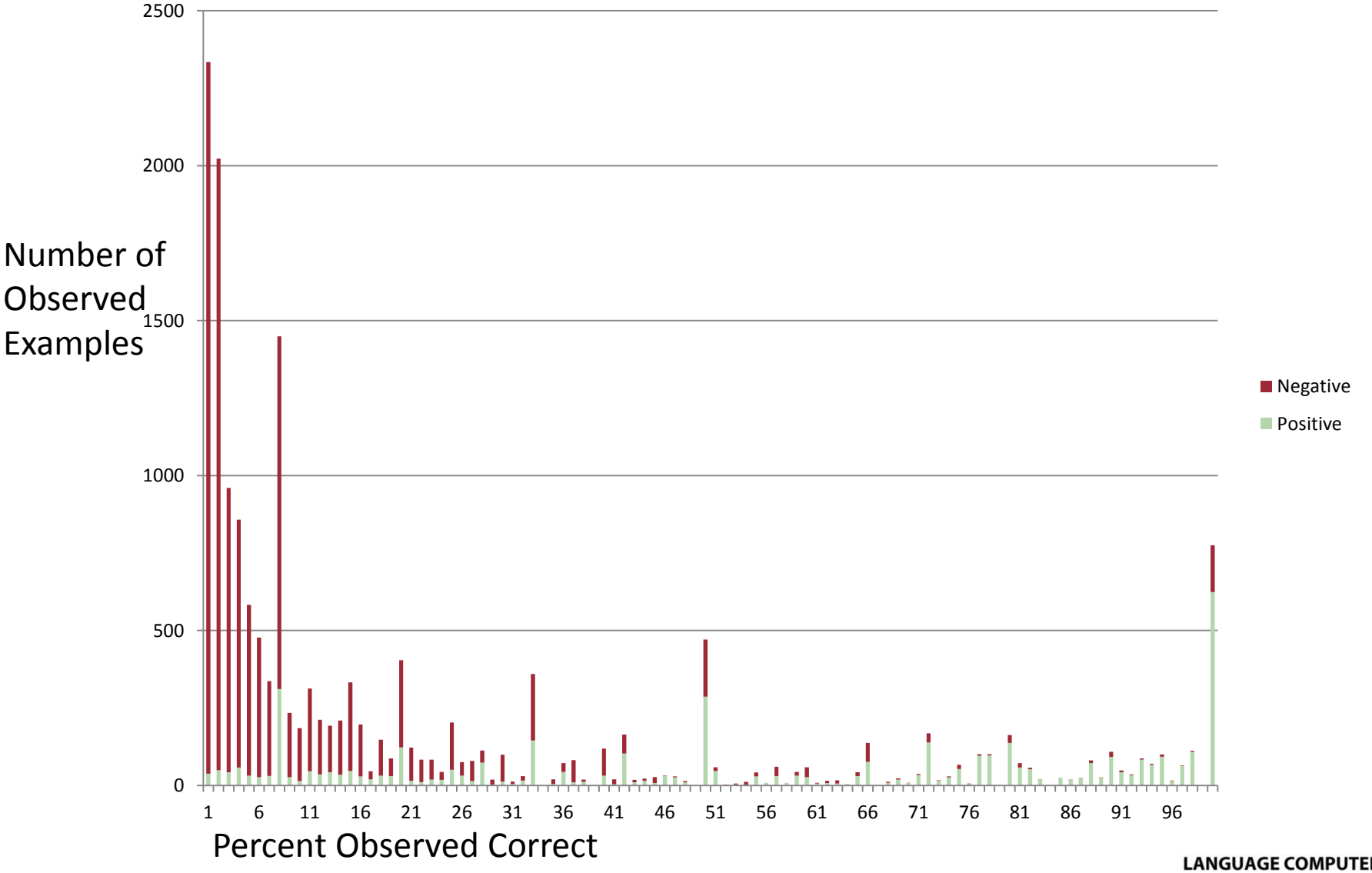
- Many of the strategies didn't work



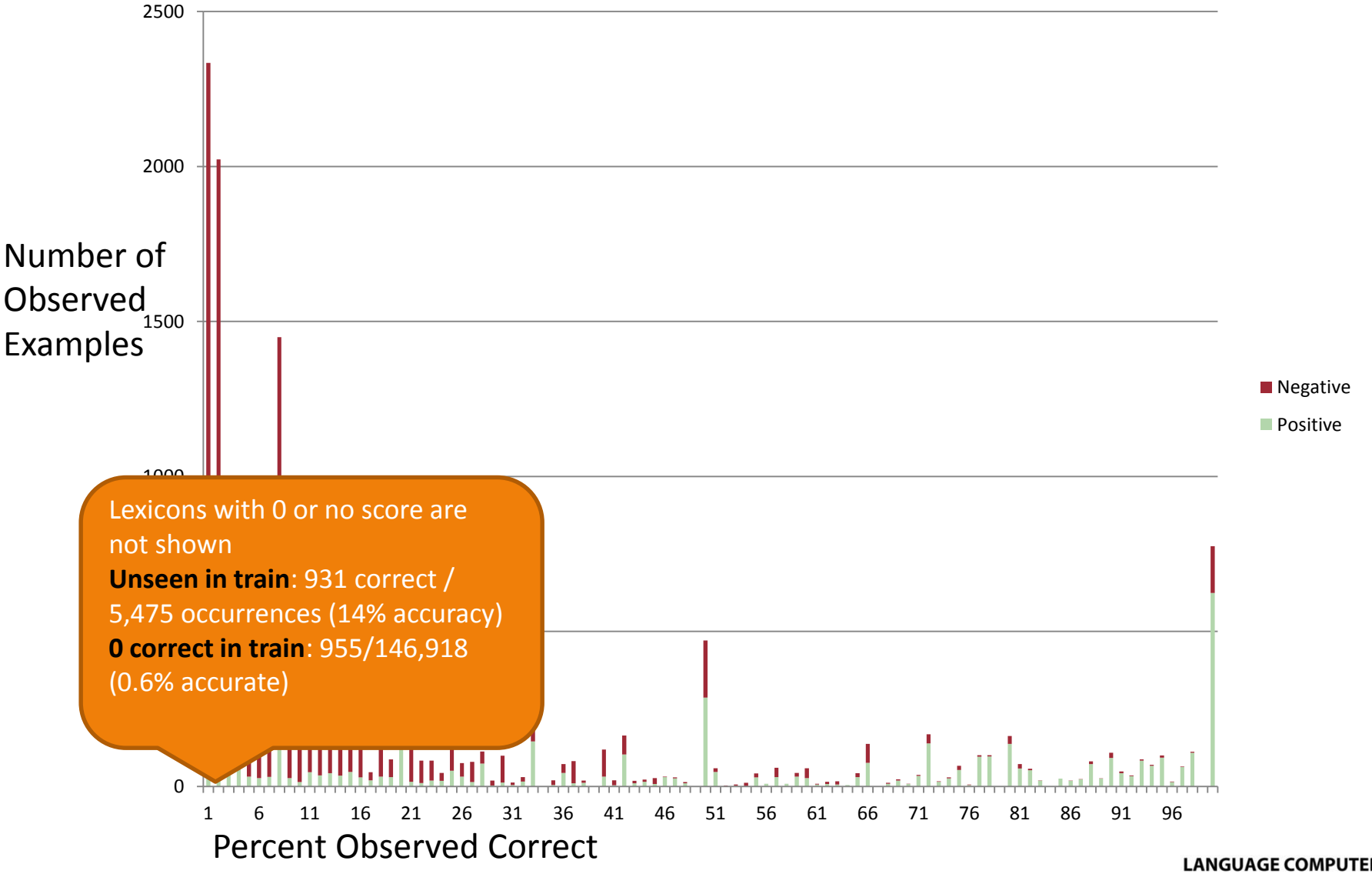
Event Detection – Lexicon Strategy

- Build a lexicon from training sources for nuggets
- C_P_word: Count the times the word/phrase occurs as a positive example
- C_T_word: Count the times the word/phrase occurs as a string
- Lexicon_score_word = C_P_word / C_T_word
- Also experimented with
  - Lexicon_score_lemma
    - Attack, attacks, attackers
  - Lexicon_score_pos
    - Attack#n, Attack#v
  - Lexicon_score_lemma_pos
    - Attacked, attacking -> Attack#v
    - Attackers, the attack -> Attack#n

Event Detection – Lexical Priors

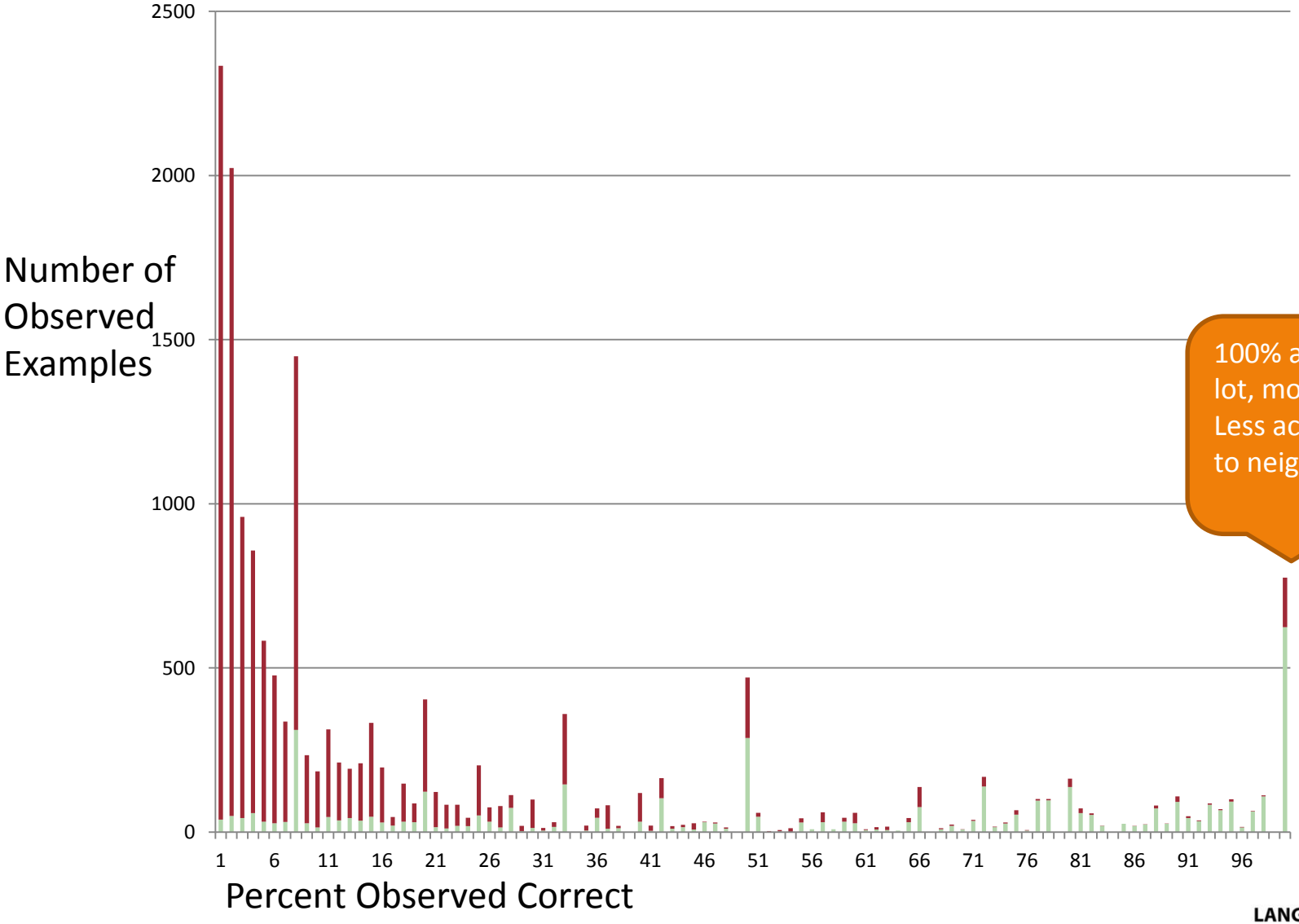


Event Detection – Lexical Priors



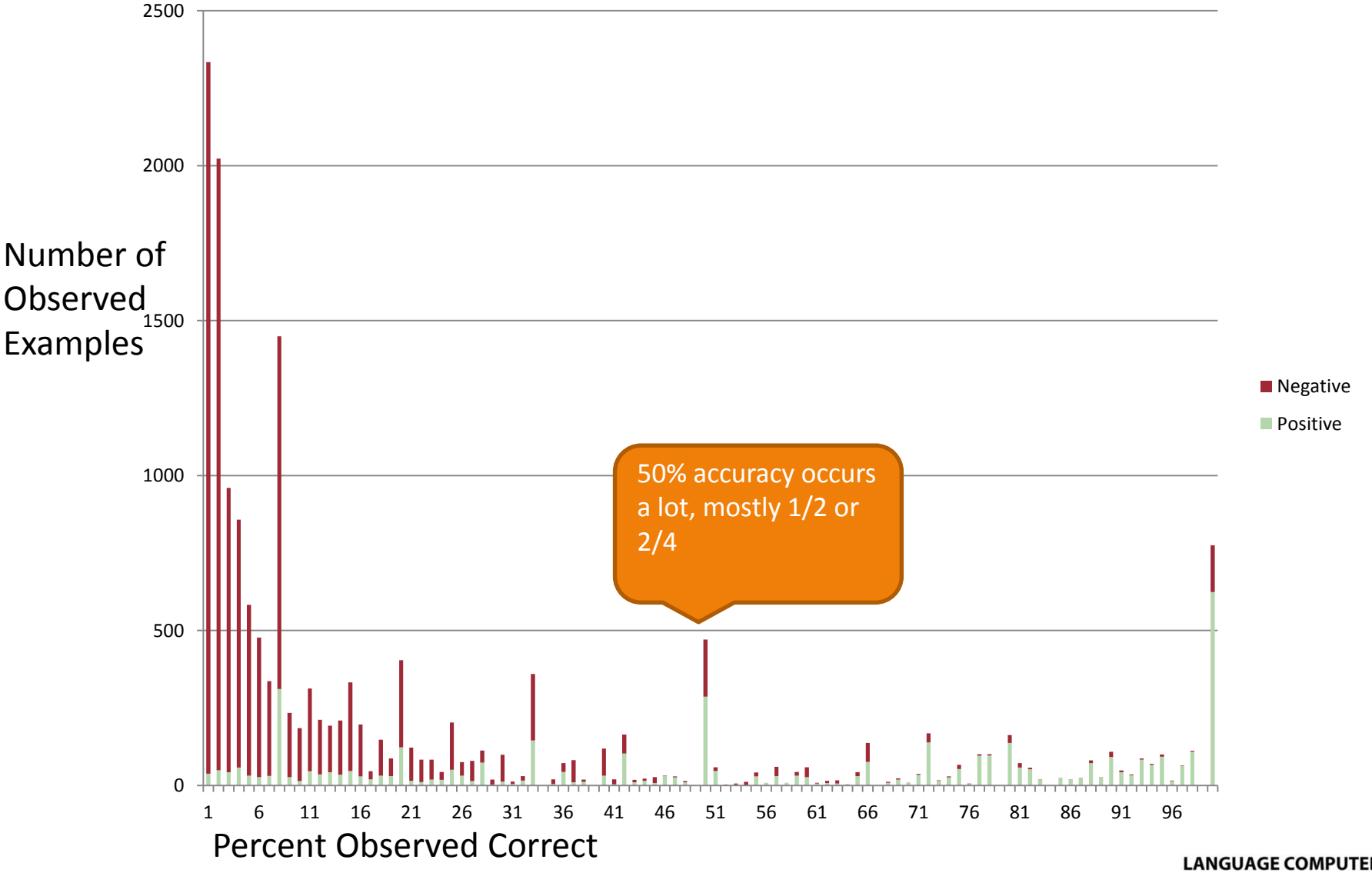
LANGUAGE COMPUTER

Event Detection – Lexical Priors



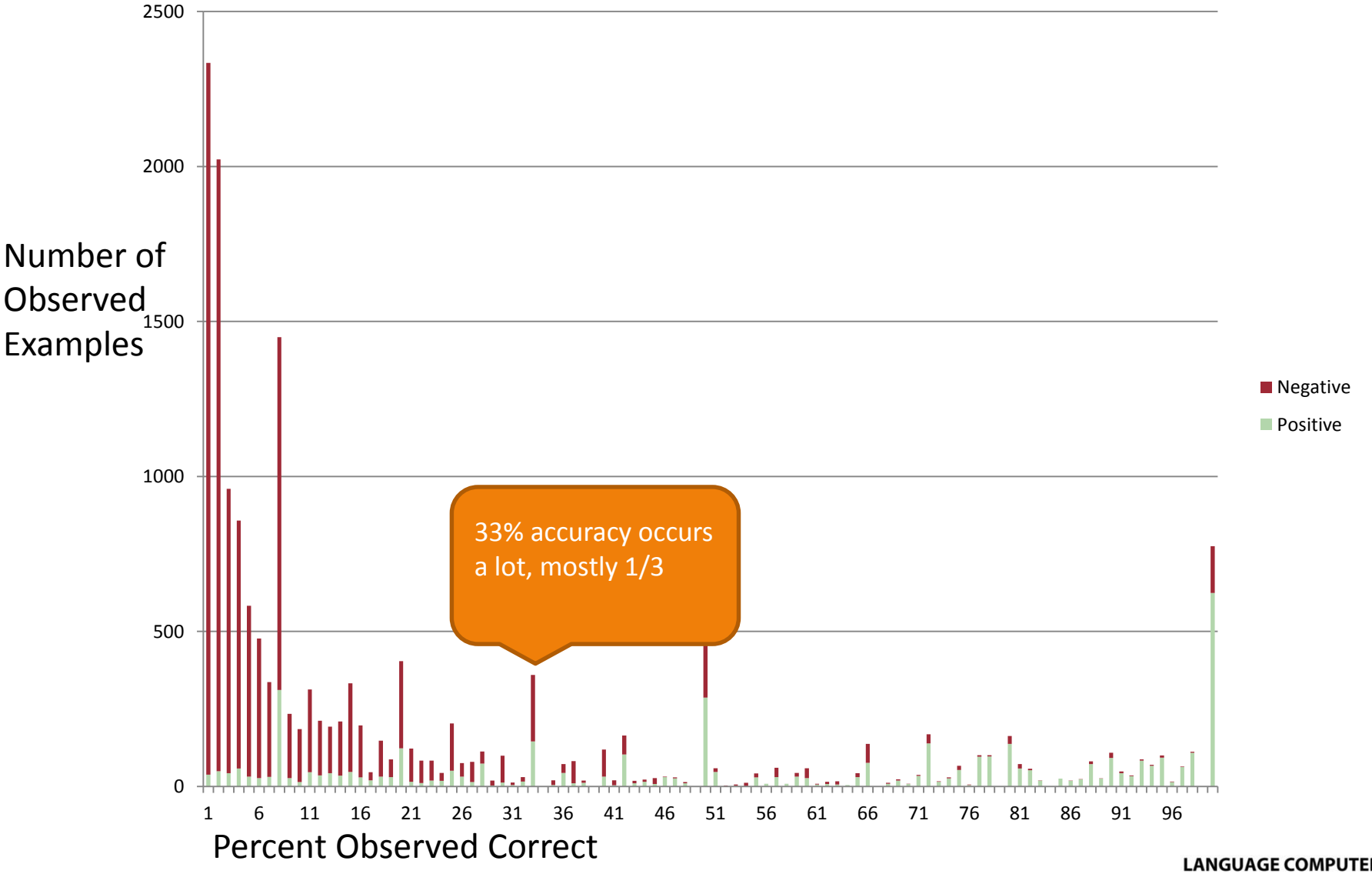
LANGUAGE COMPUTER

Event Detection – Lexical Priors



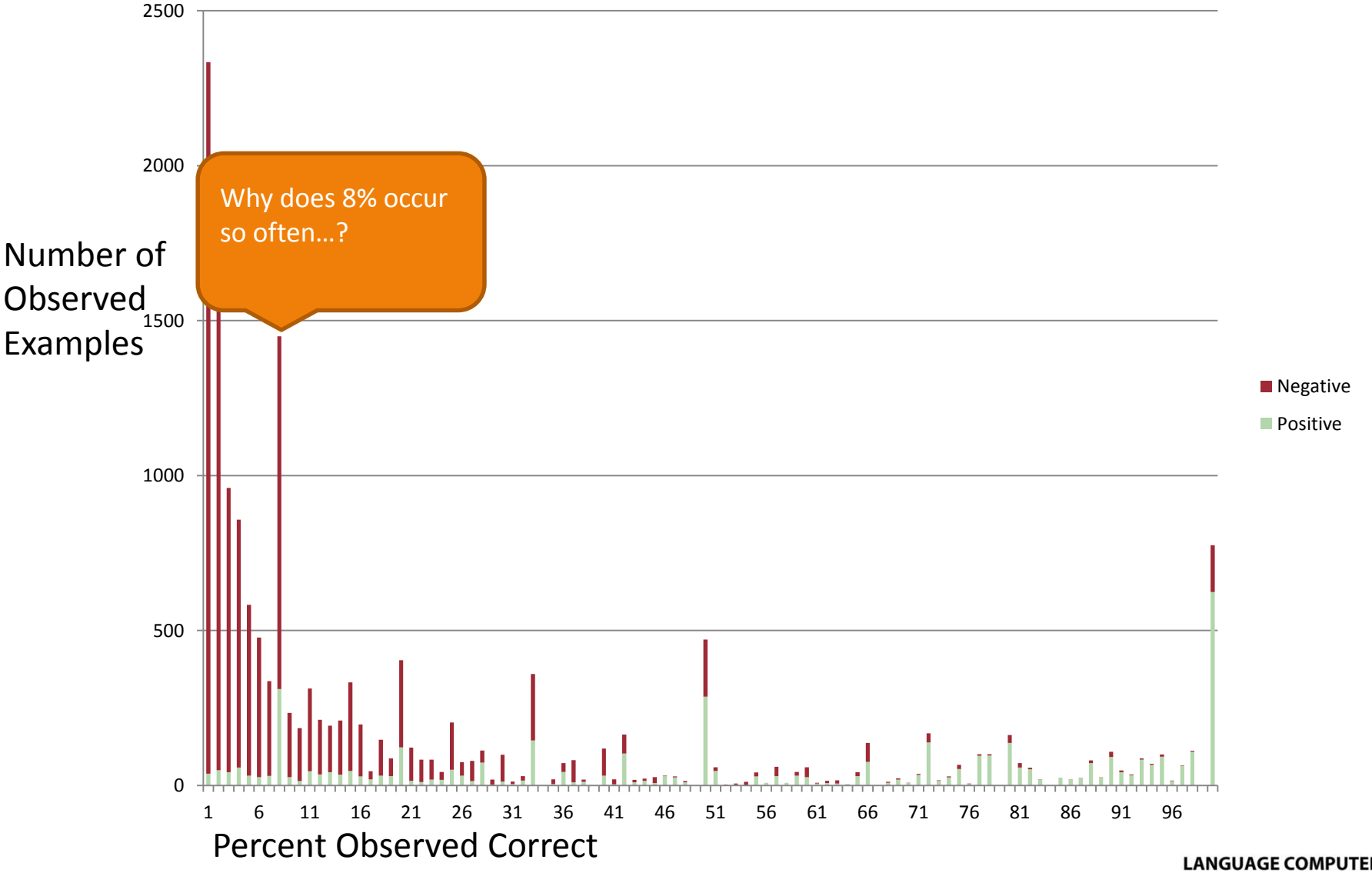
LANGUAGE COMPUTER

Event Detection – Lexical Priors



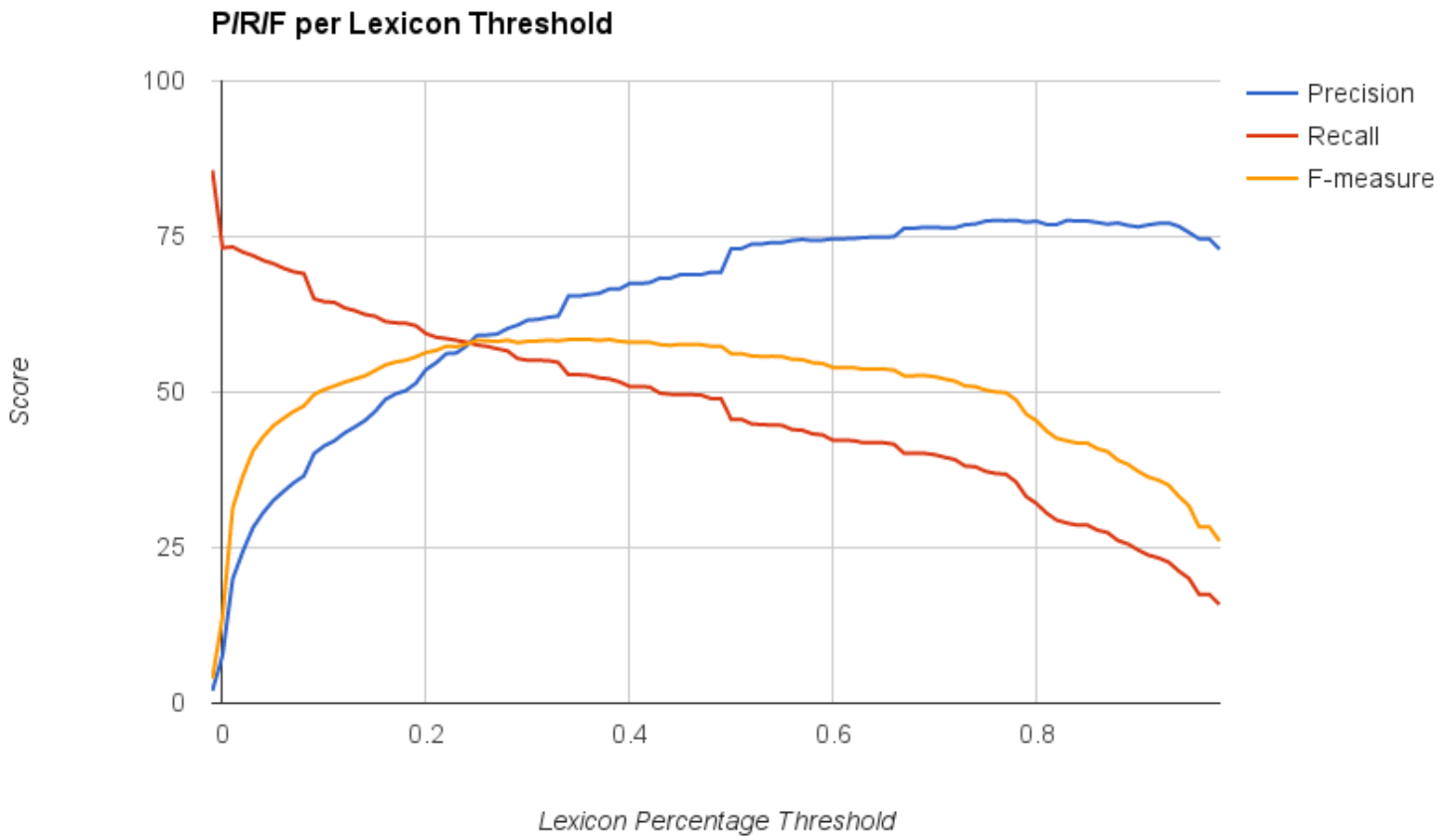
LANGUAGE COMPUTER

Event Detection – Lexical Priors



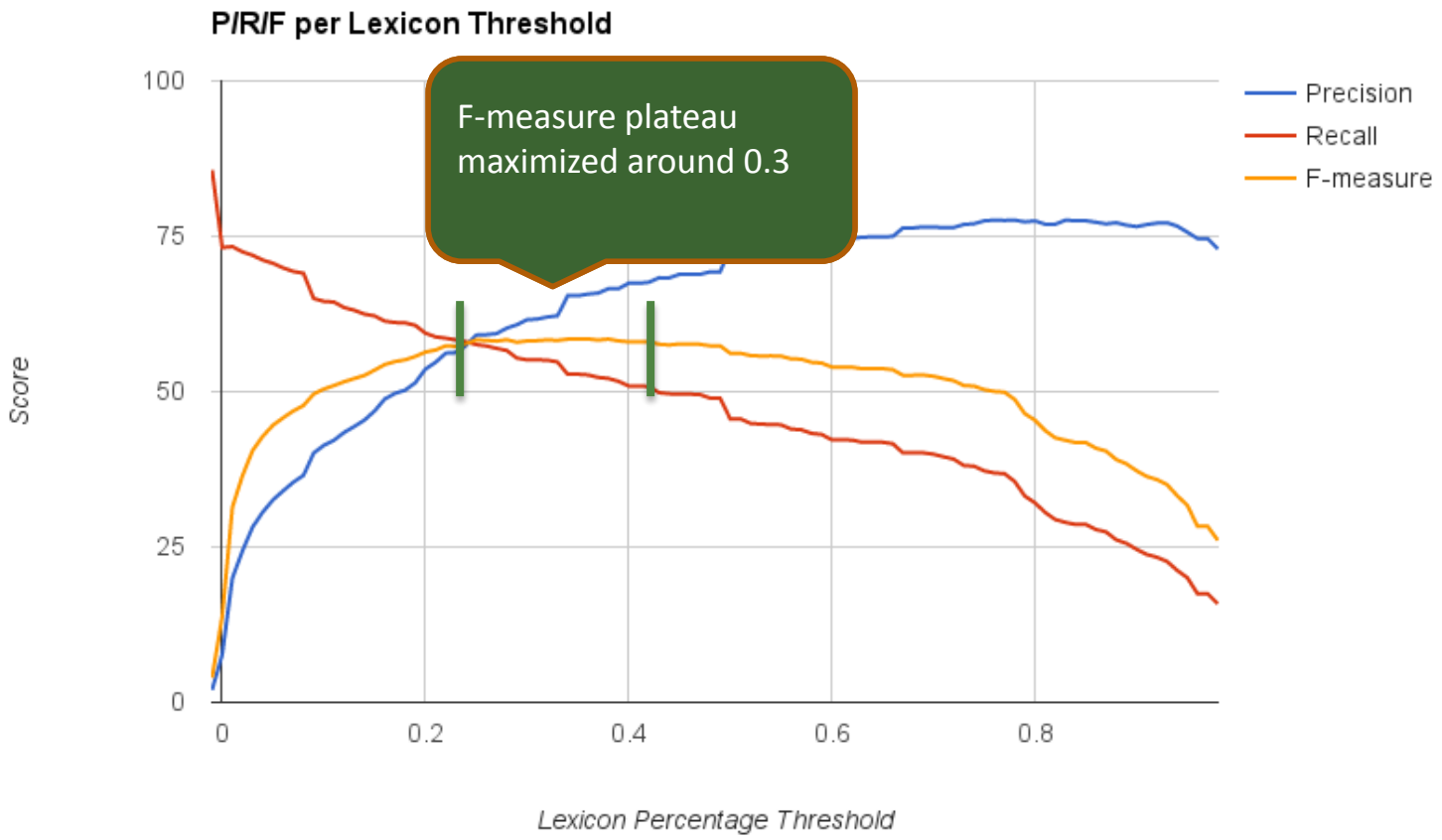
LANGUAGE COMPUTER

Event Detection – Selecting Threshold



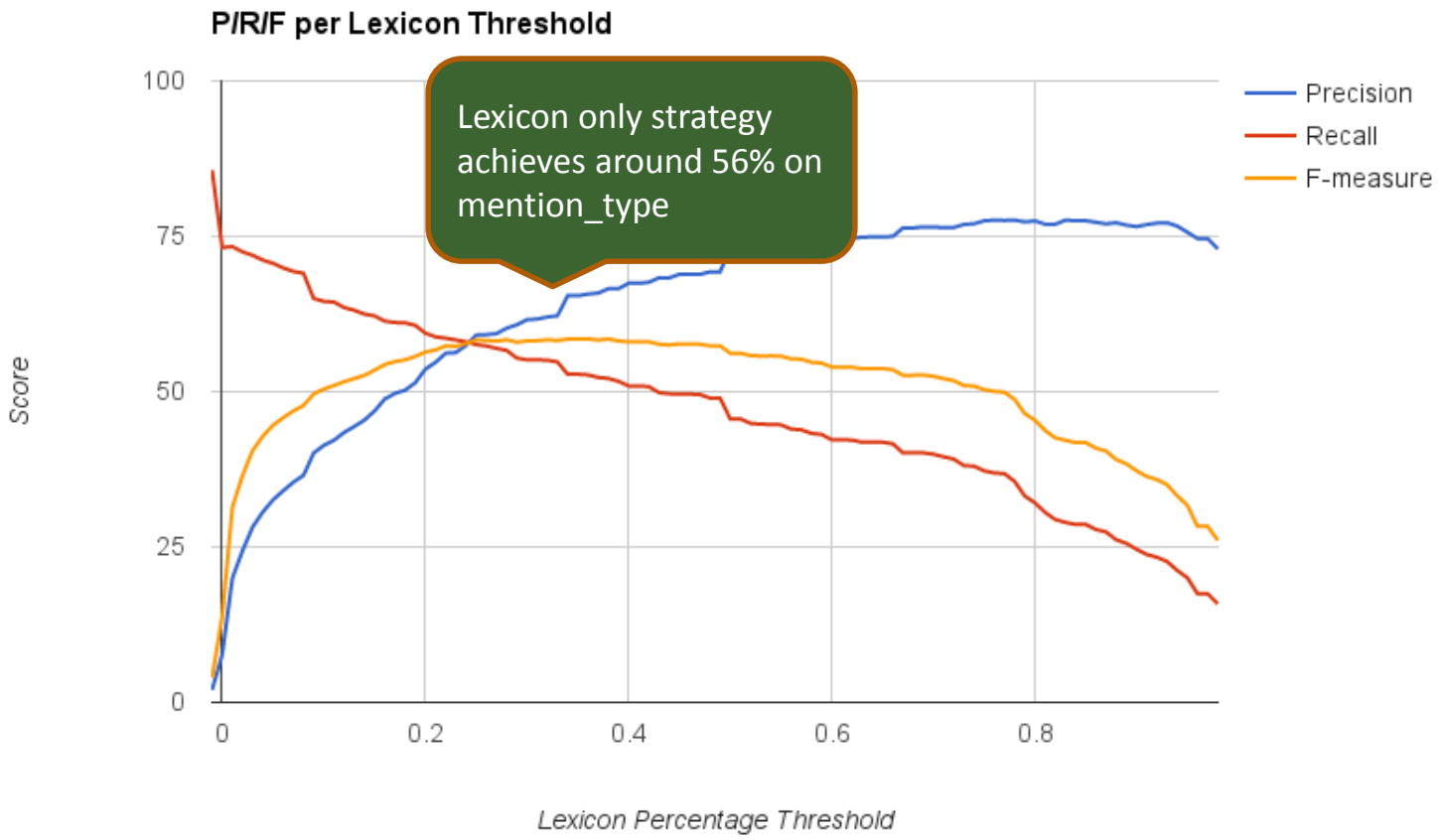
LANGUAGE COMPUTER

Event Detection – Selecting Threshold



LANGUAGE COMPUTER

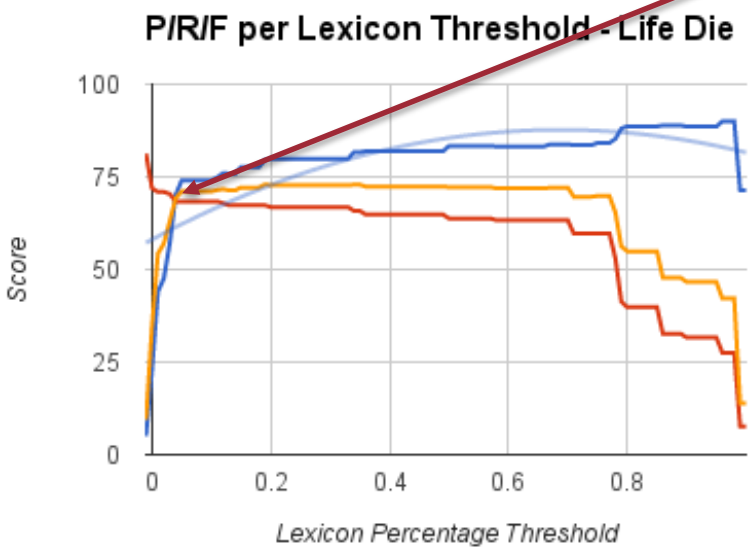
Event Detection – Selecting Threshold

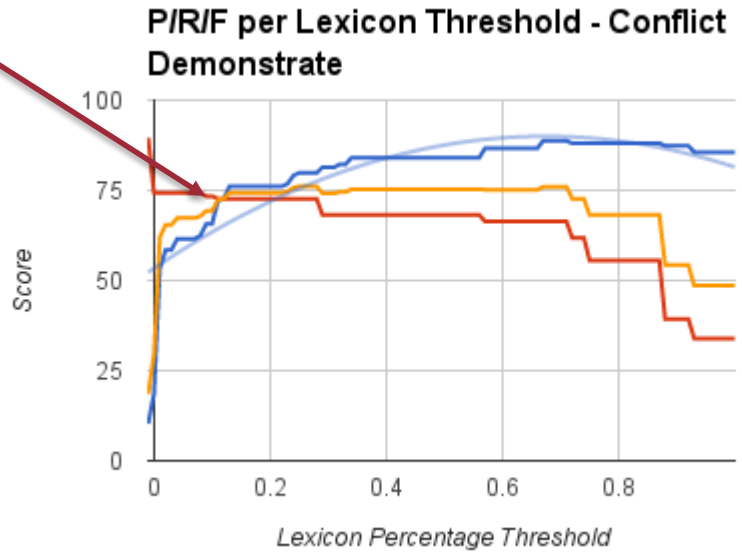


LANGUAGE COMPUTER

Event Detection – High Precision Types

Maximum F-measure achieved at low lexicon threshold



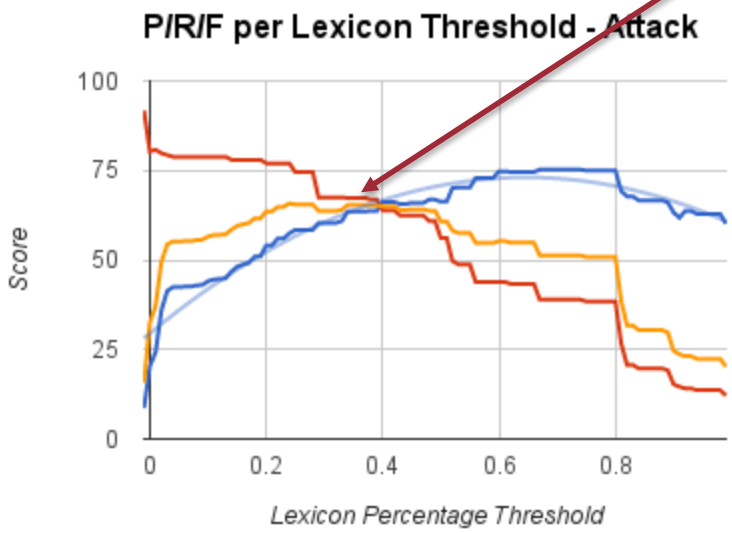


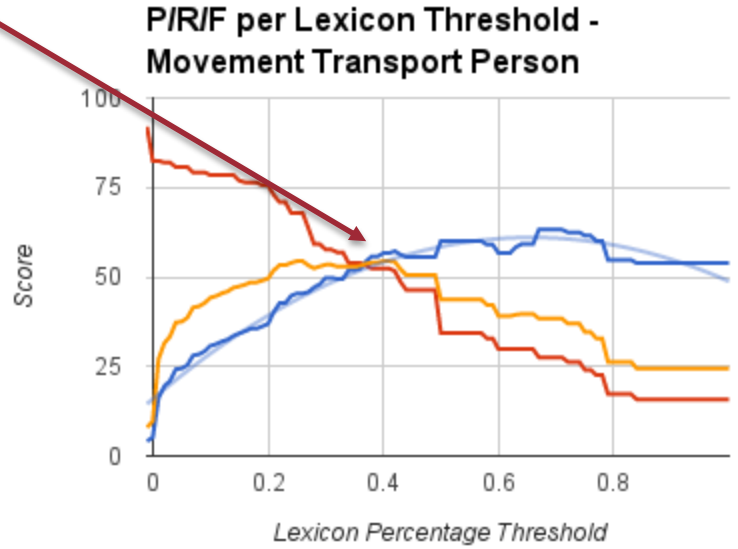
Recall
Precision
F-Measure
Precision Trendline

LANGUAGE COMPUTER

Event Detection – Medium Precision Types

Maximum F-measure achieved at higher lexicon threshold

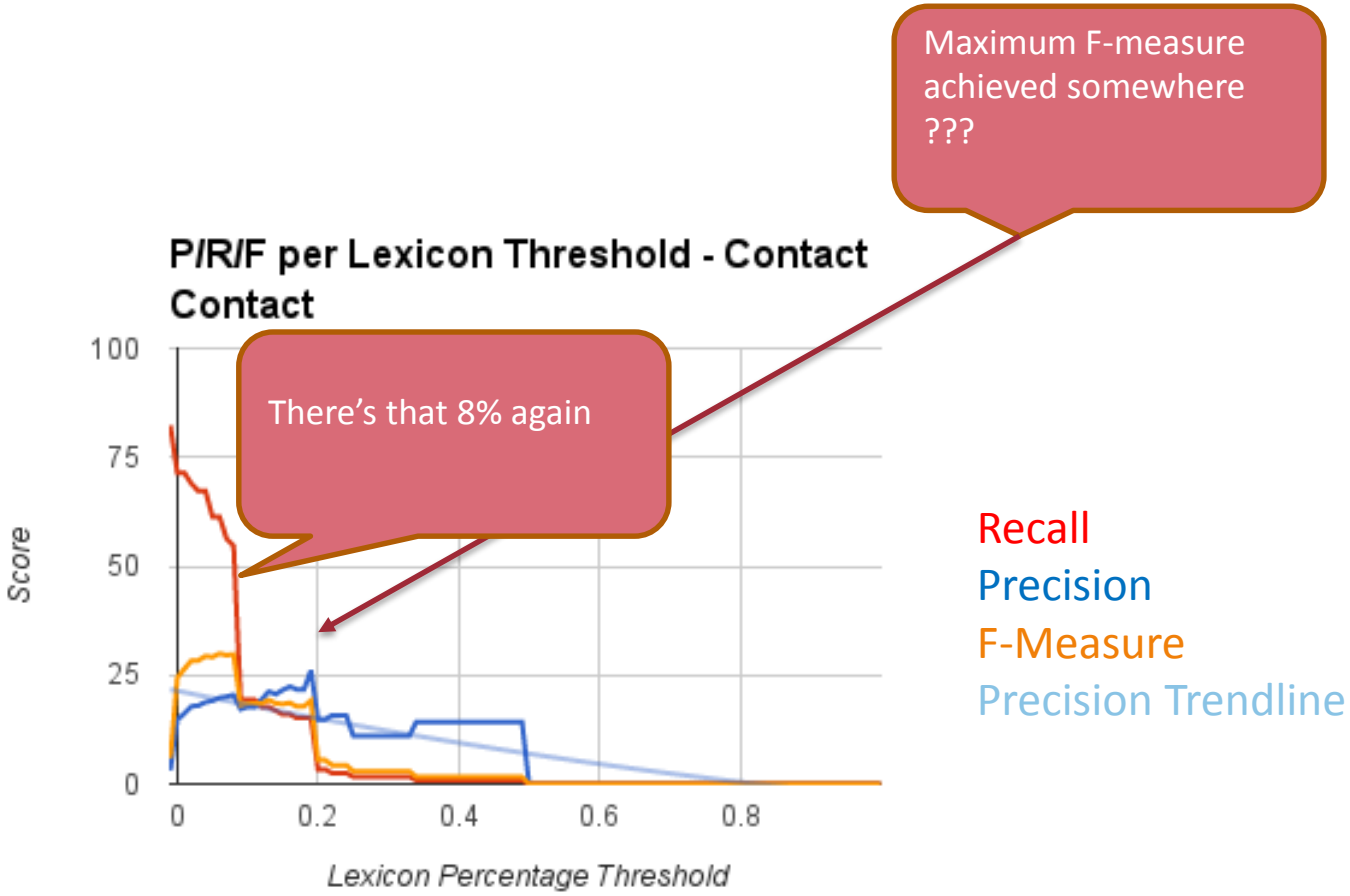




Recall
Precision
F-Measure
Precision Trendline

LANGUAGE COMPUTER

Event Detection – Low Precision Types

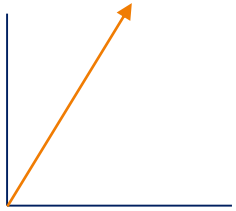


LANGUAGE COMPUTER

Event Detection – Context Modelling

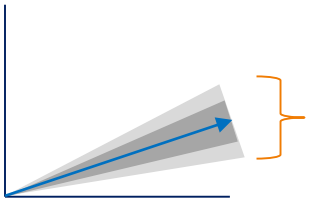
Example: Justice Sentence

John was given a life sentence.



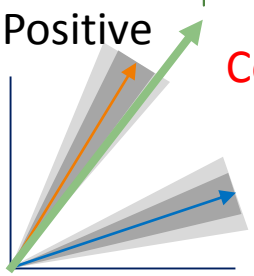
Vector representation for context
(Doc2Vec, Le and Mikolov, 2014)

John wrote a sentence about life.
The sentence had 17 words.



Estimated Density Function For Negatives

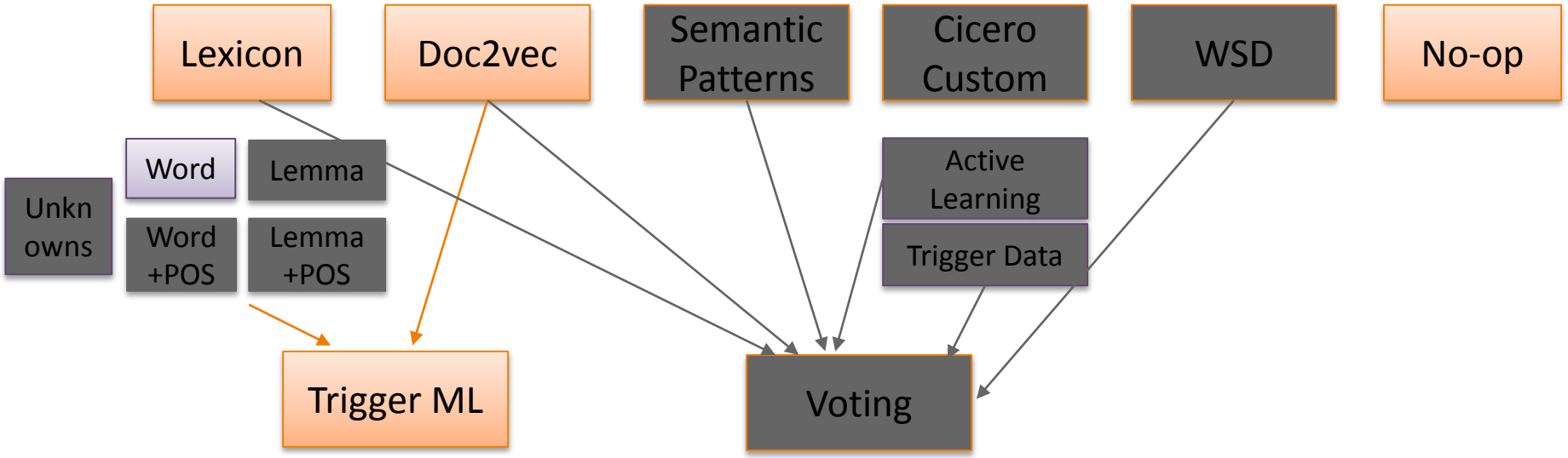
Peter's life sentence was almost over.



Contextual Classification

Event Detection – Winning Strategies

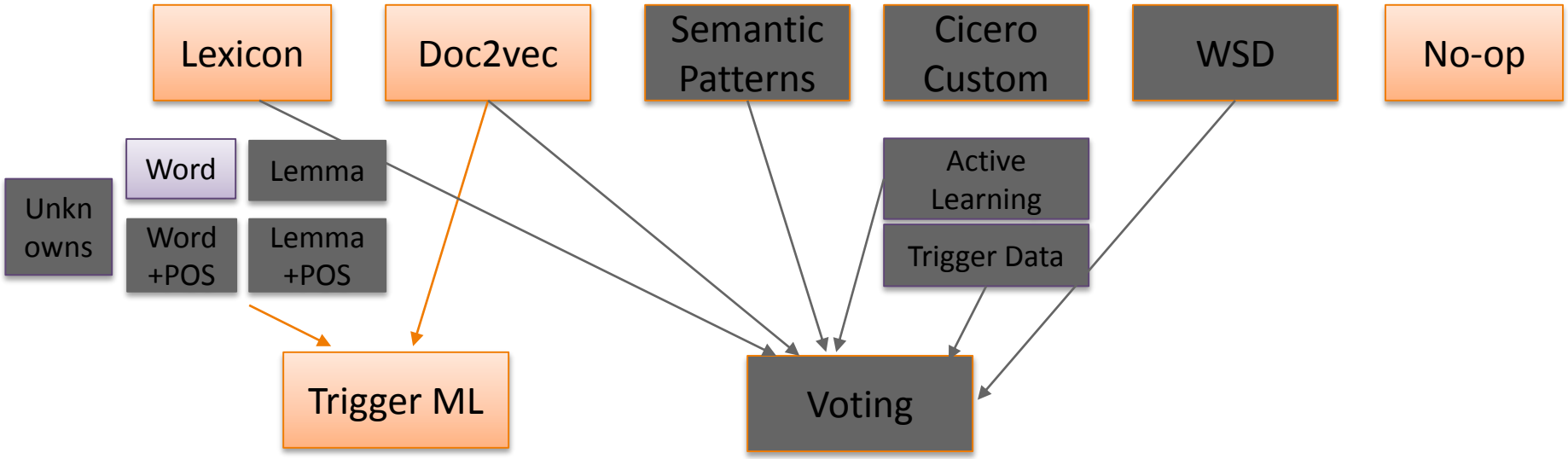
- Pick best combination of strategies for each event type
  - Watch out for Micro- vs. Macro F-measure
    - In order to optimize Micro, we use the No-op strategy for some types



Event Detection – Winning Strategies

- Pick best combination of strategies for each event type
  - Watch out for Micro- vs. Macro F-measure

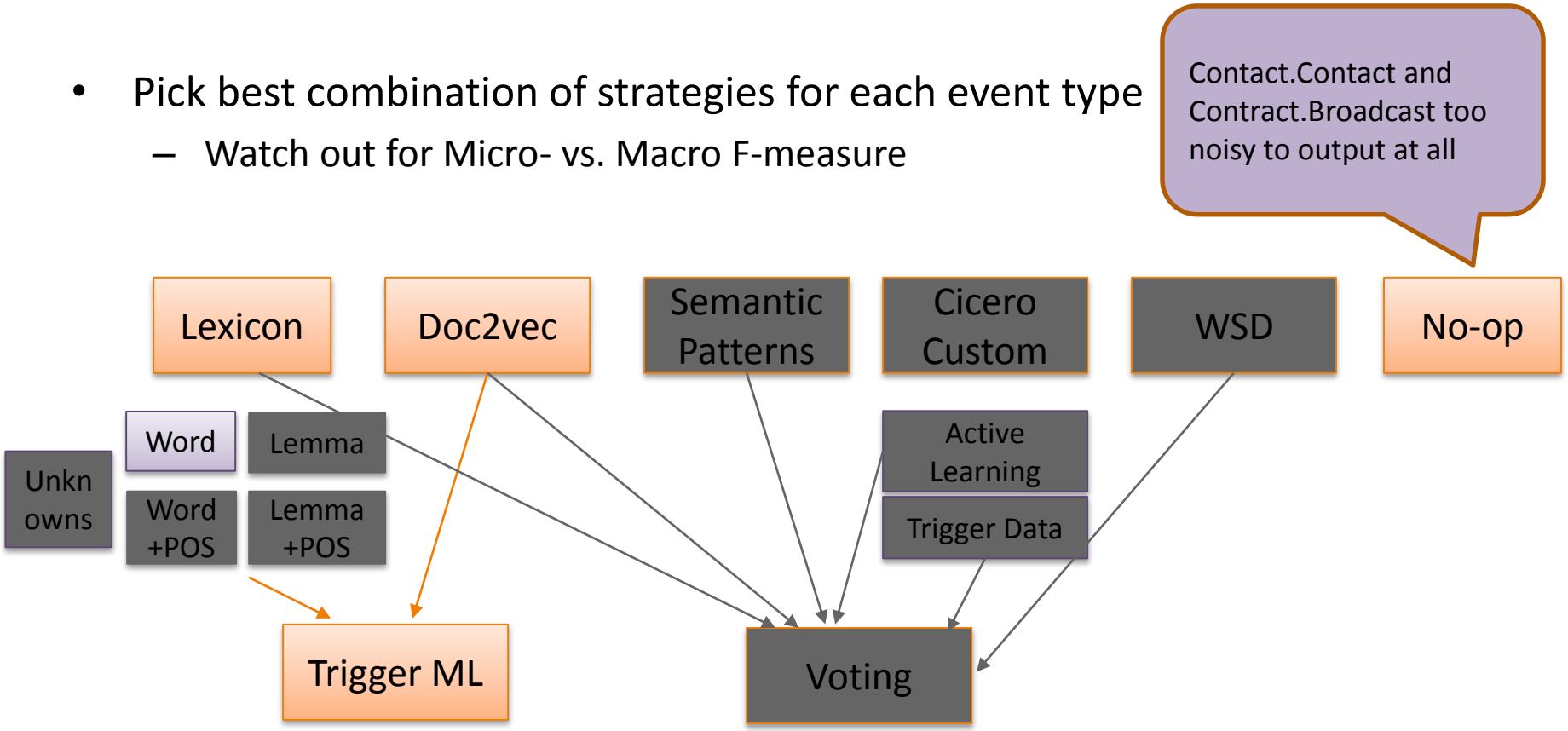
End-Org, Manufacture.Artifact, Transaction.Transaction occur too rarely to model



LANGUAGE COMPUTER

Event Detection – Winning Strategies

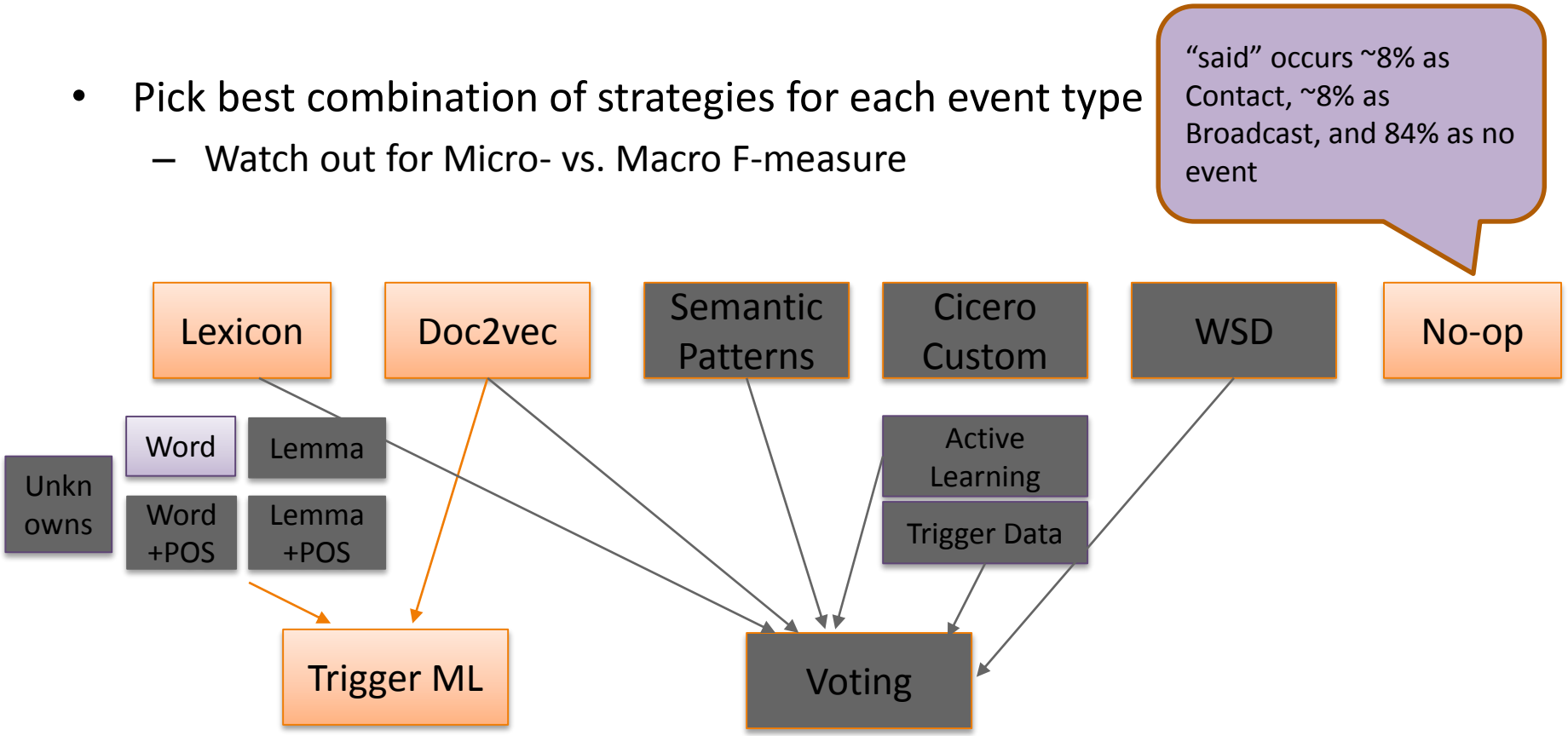
- Pick best combination of strategies for each event type
  - Watch out for Micro- vs. Macro F-measure



LANGUAGE COMPUTER

Event Detection – Winning Strategies

- Pick best combination of strategies for each event type
  - Watch out for Micro- vs. Macro F-measure



Event Detection – Evaluation

Task 1

test	Event (mention_type)			+realis_status		
	P	R	F	P	R	F
LCC1	66.86	53.31	59.32	49.80	39.71	44.18

eval	Event (mention_type)			+realis_status		
	P	R	F	P	R	F
Rank1			58.41			44.24
LCC2	73.95	45.61	57.18	49.22	31.02	38.06
LCC1	72.92	45.91	56.35	48.92	30.81	37.81
Median			48.79			34.78

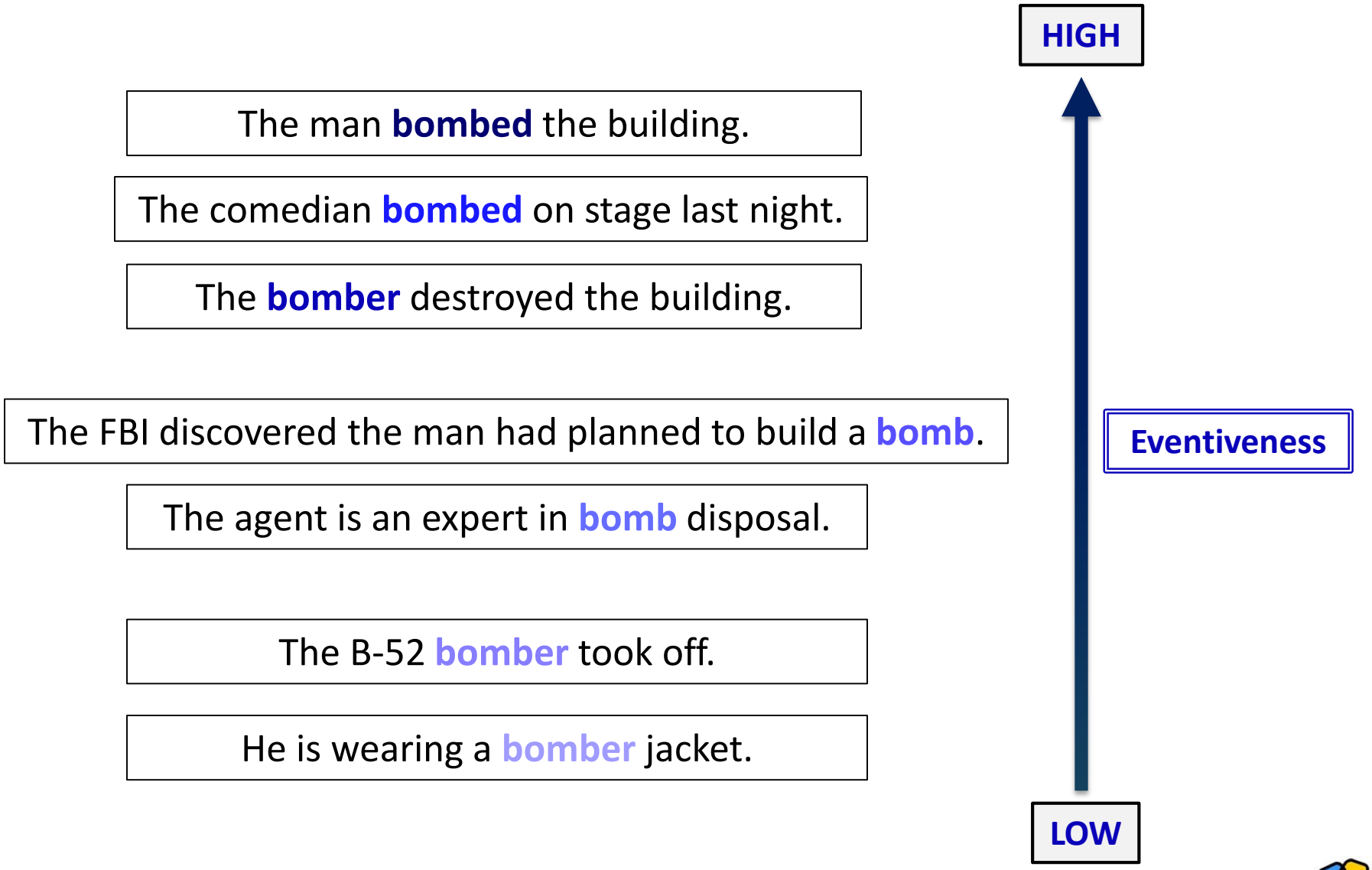
Event Detection – Challenge

- Data is one-dimensional
  - This text is a trigger for this event type

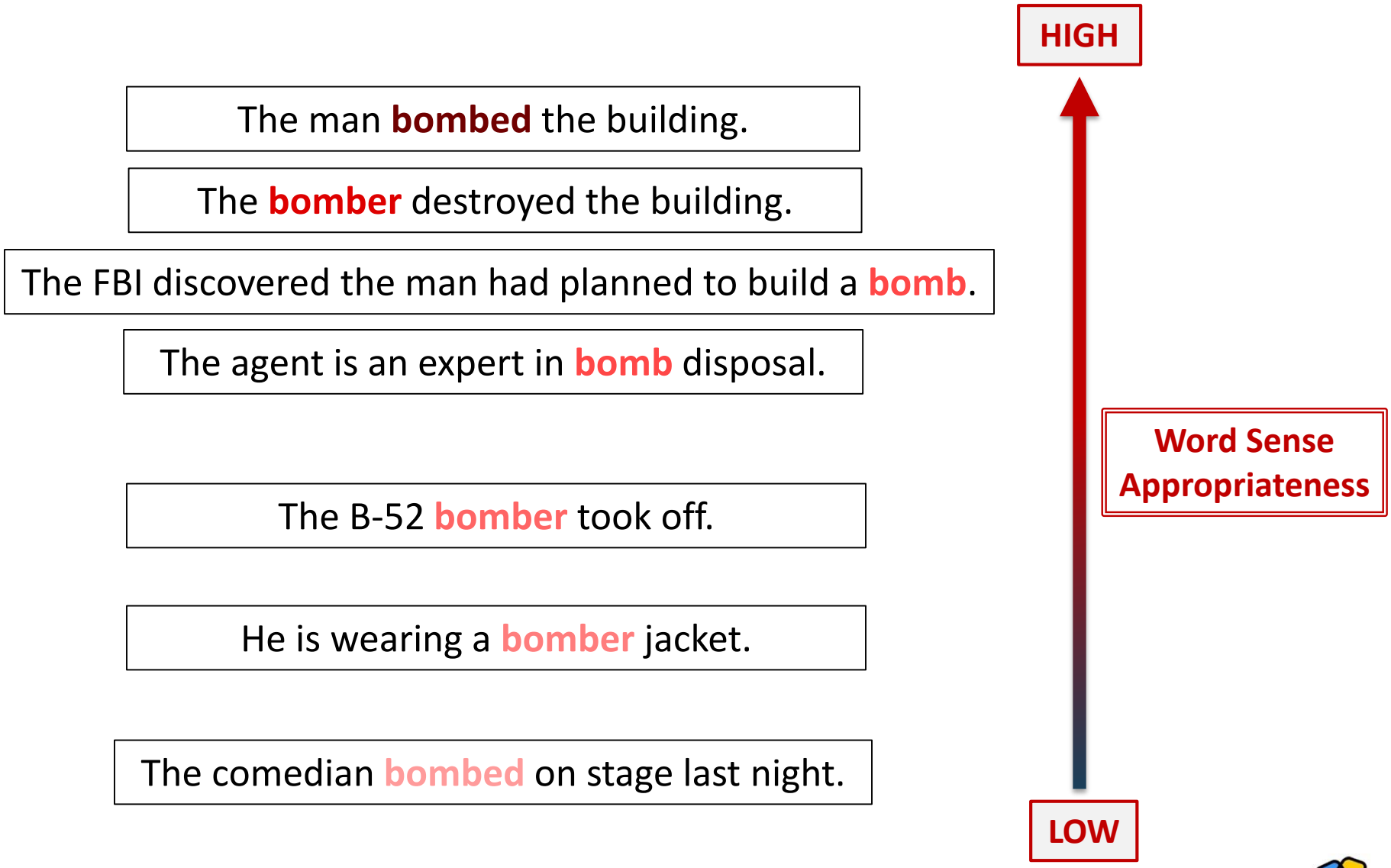
- Problem is multi-dimensional
  1. Does this meet the minimum threshold to be considered an “event”?
  2. Is this text describing the appropriate event type?

- Could access to extra annotation data provide a solution?

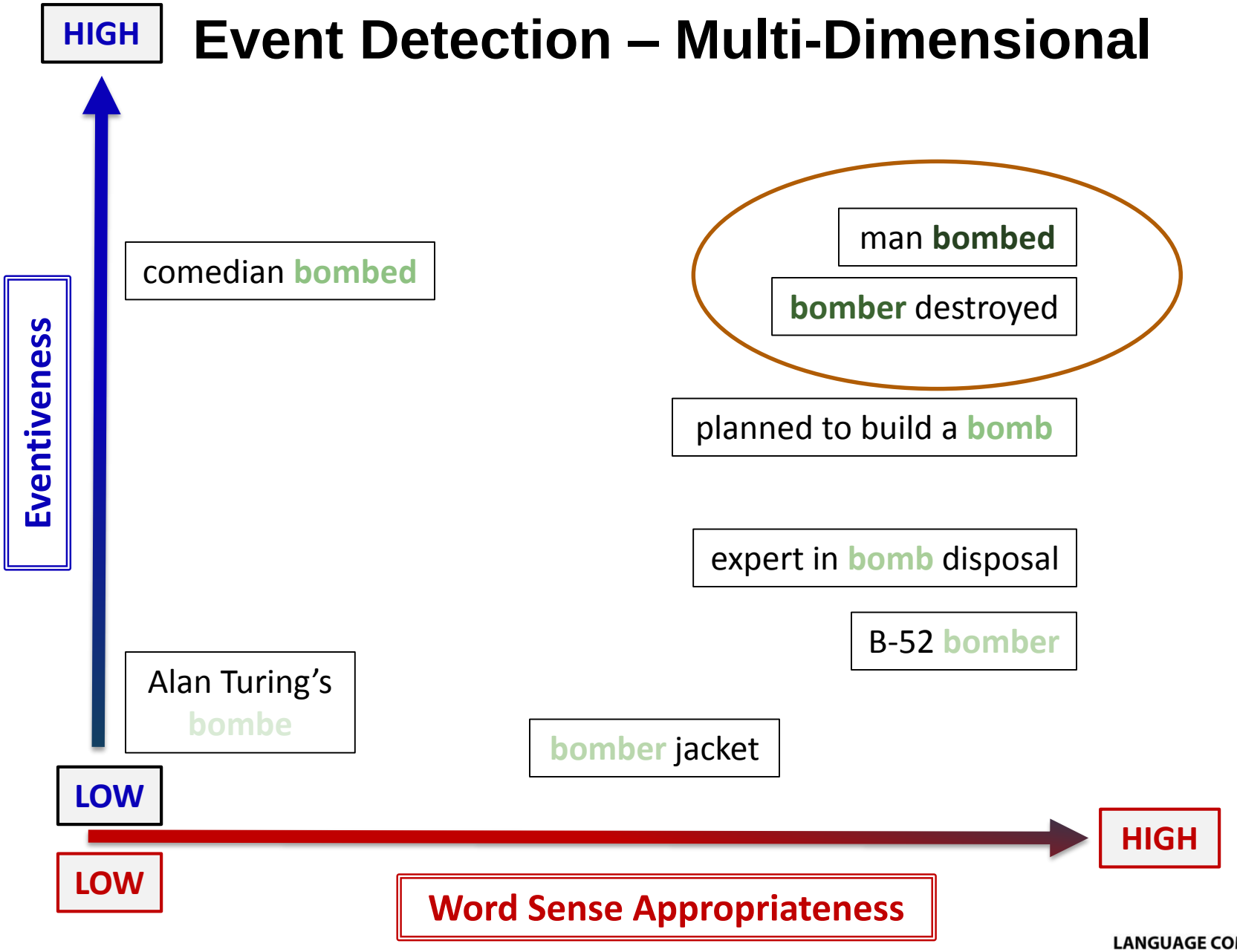
Event Detection – Eventiveness



Event Detection – Word Sense Appropriateness



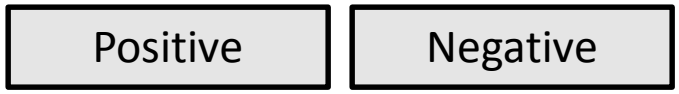
Event Detection – Multi-Dimensional



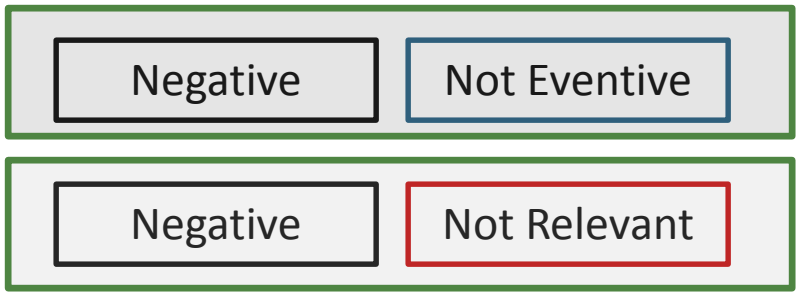
LANGUAGE COMPUTER

Event Detection – Detailed Annotations

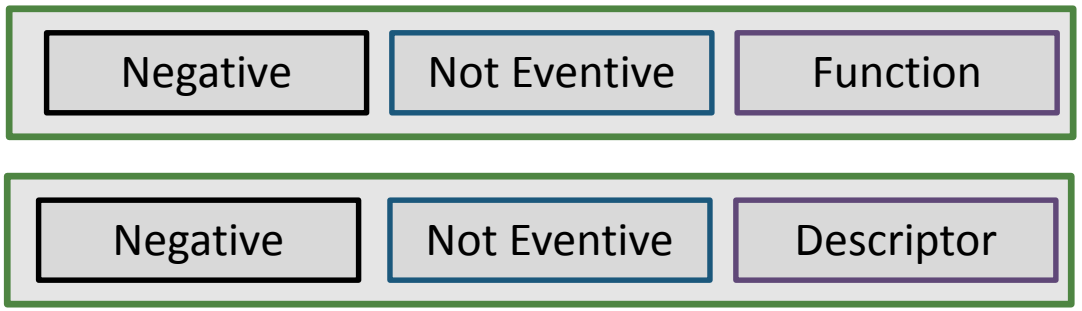
1. **One-dimensional** outcome



2. **Two-dimensional** outcome



3. **Three-dimensional** outcome
   - *B52-bomber*
   - *Abusive Husband*



LANGUAGE COMPUTER

Overview

- Event Detection (Task 1)
- Event Hoppers (Task 2 / 3)
  - Compatibility Modules
  - Hopperator
  - Scores on Diagnostic vs. System events

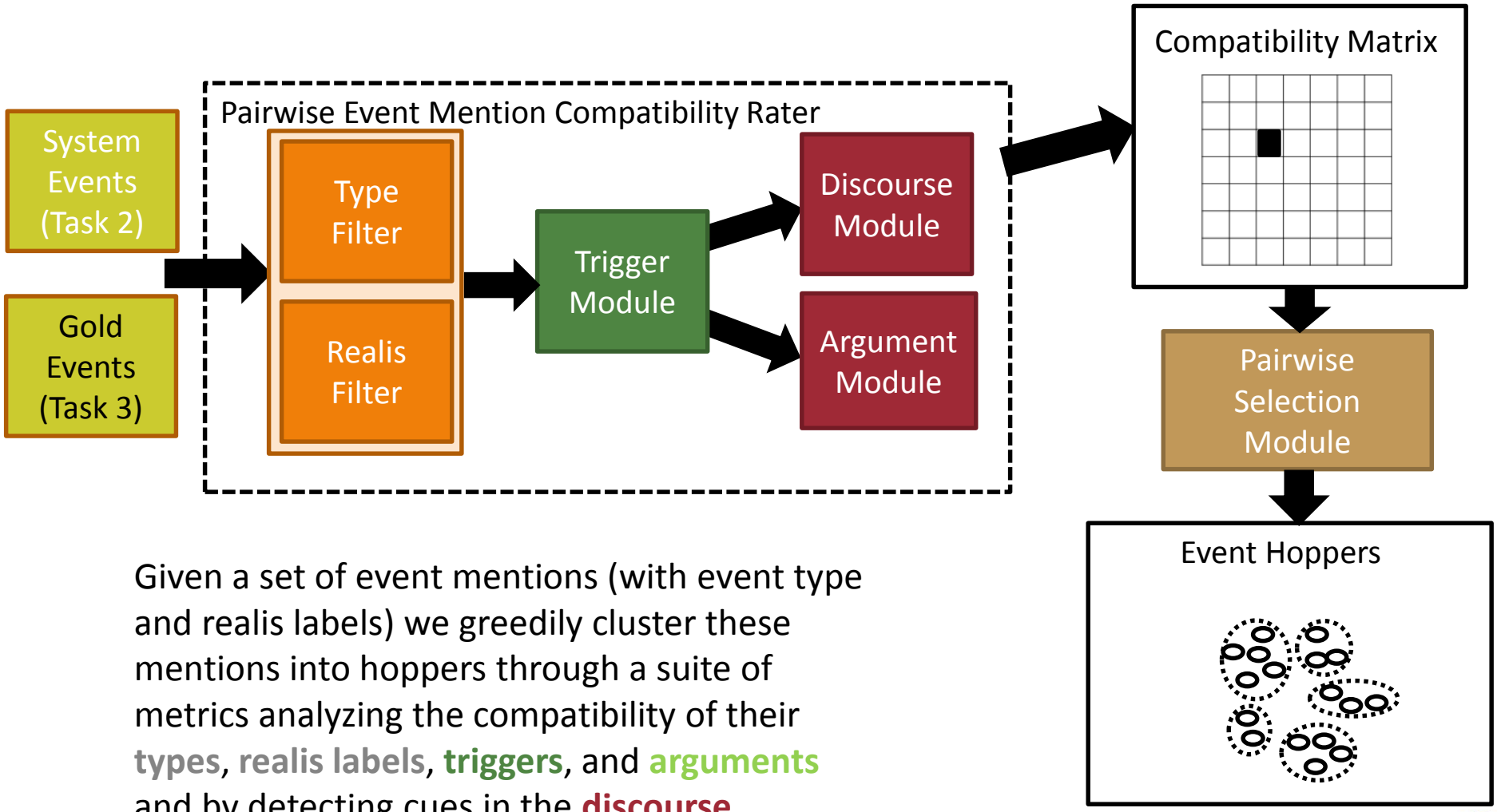
Event Hoppers - Description

- Event Hoppers consist of event mentions that refer to the same event occurrence.
- For this purpose, we define a more inclusive, less strict notion of event coreference as compared to ACE and Light ERE.
- Event hoppers contain mentions of events that **"feel" coreferential** to the annotator.
- Event mentions that have the following features go into the same hopper:
  - They have the **same event type** and **subtype** (with exceptions for Contact.Contact and Transaction.Transaction)
  - They have the **same temporal** and **location scope**.
- The following do not represent an incompatibility between two events.
  - **Trigger specificity** can be **different** (assaulting 32 people vs. wielded a knife)
  - Event **arguments** may be **non-coreferential** or **conflicting** (18 killed vs. dozens killed)
  - **Realis status** may be **different** (will travel [OTHER] to Europe next week vs. is on a 5-day trip [ACTUAL])

Event Hoppers – Metrics

- Formal
  - KBP – the arithmetic mean of the following four metrics for clustering evaluation:
  - B-Cubed, MUC, CEAFE, and BLANC.
  - Note: A script was provided by the KBP organizers to run these four metrics and compute the mean.
- Internal Metrics
  - Provides a way to compare systems that the formal metric does not
  - PairP – hopper precision over event mention pairs (PairP = JNT/SH)
  - PairR – hopper recall over event mention pairs (PairR = JNT/GH)
  - GH is the number of event mention pairs in the gold-standard hoppers
  - SH is the number of pairs in the system-generated hoppers
  - JNT is the number of system hopper pairs that are also paired in the gold hoppers

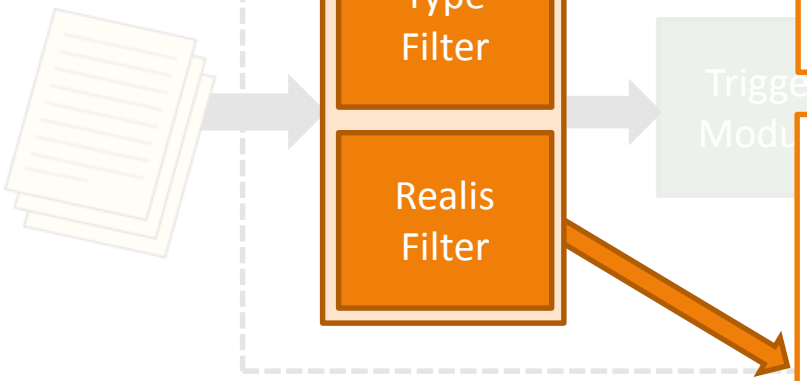
Event Hoppers – Faceted Approach



Given a set of event mentions (with event type and realis labels) we greedily cluster these mentions into hoppers through a suite of metrics analyzing the compatibility of their **types**, **realis labels**, **triggers**, and **arguments** and by detecting cues in the **discourse**.

LANGUAGE COMPUTER

Event Hoppers – Faceted Approach



1) Ensures that event pairs have compatible event types. Note that CONTACT_CONTACT and TRANSACTION_TRANSACTION are compatible with all CONTACT and TRANSACTION types respectively.
2) Ensures that event triggers do not have overlapping spans.

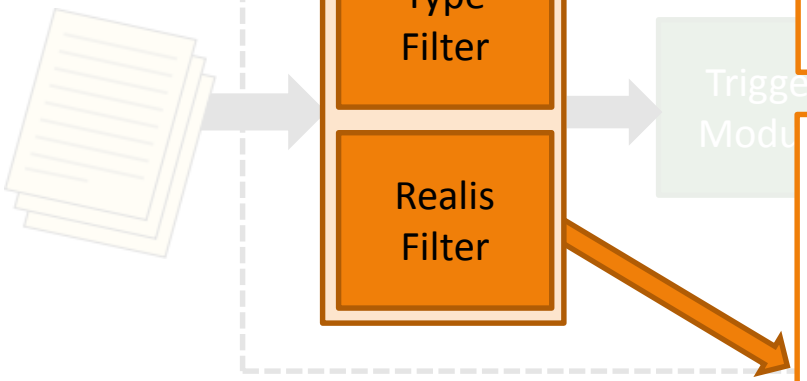
Three modes:

1. Realis is ignored.
2. GENERIC realis is incompatible with ACTUAL or OTHER. [BASIC]
3. GENERIC, ACTUAL, and OTHER are incompatible with one another, excluding ACTUAL + OTHER (future tense). [STRICT]

Method	PairP	PairR	CoNLL Score
All Singletons	0.00	0.00	37.96
All Events w/ same type	15.6	**65.6**	46.65
R=BASIC	19.3	65.0	48.65
R=STRICT	**23.6**	63.3	**50.69**

Task 2

Event Hoppers – Faceted Approach



1) Ensures that event pairs have compatible event types. Note that CONTACT_CONTACT and TRANSACTION_TRANSACTION are compatible with all CONTACT and TRANSACTION types respectively.
2) Ensures that event triggers do not have overlapping spans.

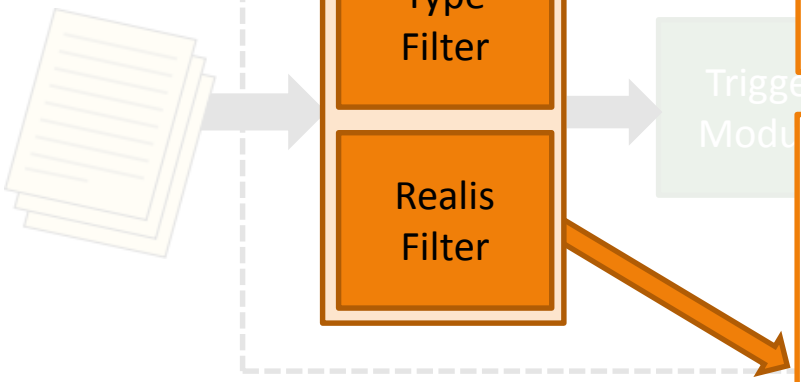
Three modes:

1. Realis is ignored.
2. GENERIC realis is incompatible with ACTUAL or OTHER. [BASIC]
3. GENERIC, ACTUAL, and OTHER are incompatible with one another, excluding ACTUAL + OTHER (future tense). [STRICT]

Method	PairP	PairR	CoNLL Score
All Singletons	0.00	0.00	48.85
All Events w/ same type	18.3	**99.1**	57.52
R=BASIC	22.9	94.7	61.24
R=STRICT	**30.4**	88.8	**66.30**

Task 3

Event Hoppers – Faceted Approach



1) Ensures that event pairs have compatible event types. Note that CONTACT_CONTACT and TRANSACTION_TRANSACTION are compatible with all CONTACT and TRANSACTION types respectively.
2) Ensures that event triggers do not have overlapping spans.

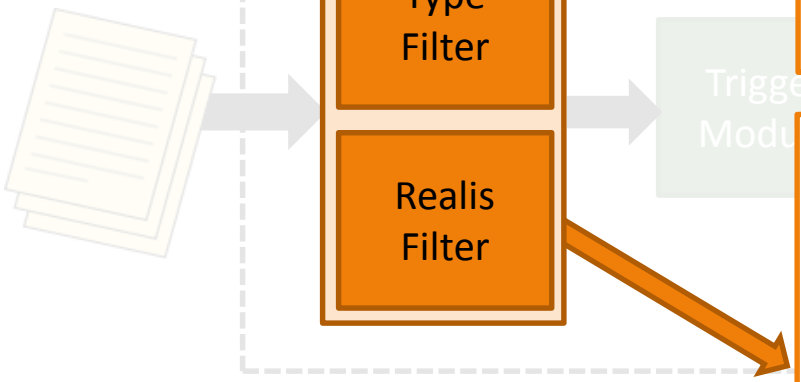
Three modes:

1. Realis is ignored.
2. GENERIC realis is incompatible with ACTUAL or OTHER. [BASIC]
3. GENERIC, ACTUAL, and OTHER are incompatible with one another, excluding ACTUAL + OTHER (future tense). [STRICT]

4.4% of pairings are ACTUAL/GENERIC

Method			CoNLL Score
All Singleton			48.85
All Events w/ same type	18.3	**99.1**	57.52
R=BASIC	22.9	94.7	61.24
R=STRICT	**30.4**	88.8	**66.30**

Event Hoppers – Faceted Approach



1) Ensures that event pairs have compatible event types. Note that CONTACT_CONTACT and TRANSACTION_TRANSACTION are compatible with all CONTACT and TRANSACTION types respectively.
2) Ensures that event triggers do not have overlapping spans.

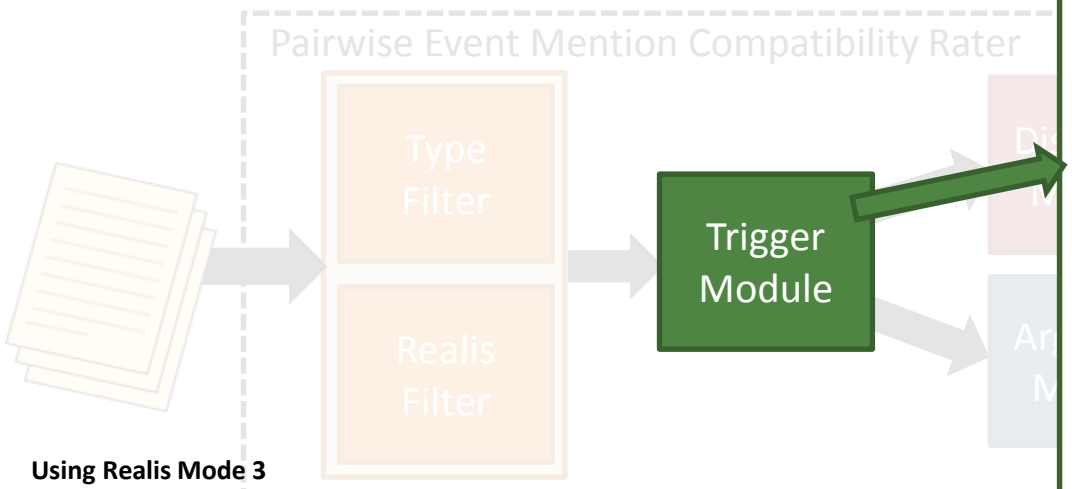
Three modes:

1. Realis is ignored.
2. GENERIC realis is incompatible with ACTUAL or OTHER. [BASIC]
3. GENERIC, ACTUAL, and OTHER are incompatible with one another, excluding ACTUAL + OTHER (future tense). [STRICT]

Method	PairP	PairR	C S
All Singletons	0.00	0.00	4
All Events w/ same type	18.3	**99.1**	
R=BASIC	22.9	94.7	61.24
R=STRICT	**30.4**	88.8	**66.30**

5.9% of pairings are ACTUAL/OTHER (excluding future tense)

Event Hoppers – Faceted Approach



Six modes:

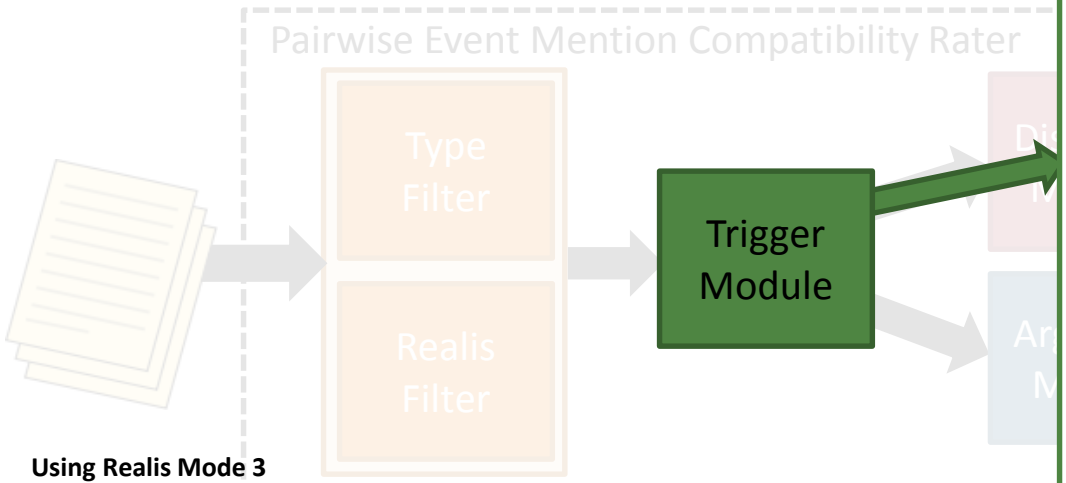
Triggers are compatible

1. …only if they match exactly.
   kills⇔kills **[EXACT]**
2. …if they share a stem.
   indicted ⇔ indicts **[SAME_STEM]**
3. …also if they share a WordNet synset or derived relationship.
   transport ⇔ ship **[SYNONYM]**
   bombings ⇔ bombed
4. …also if they can be linked by a WordNet hypernym relation.
   executed ⇔ hanged **[HYP*NYM]**
5. …also if they are included in a whitelist derived from training.
   death ⇔ fatally **[MANUAL]**
6. …for all pairs of triggers.
   shoot ⇔ impale

Using Realis Mode 3

Method	PairP	PairR	CoNLL Score
All Singletons	0.00	0.00	37.96
T=EXACT	**39.6**	27.0	54.72
T=SAME_STEM	35.6	34.7	55.69
T=SYNONYM	35.3	38.2	**56.59**
T=HYP*NYM	31.7	40.0	56.42
T=MANUAL	27.1	58.2	55.44
All Triggers Compatible	23.6	**63.3**	50.69

Event Hoppers – Faceted Approach



Six modes:

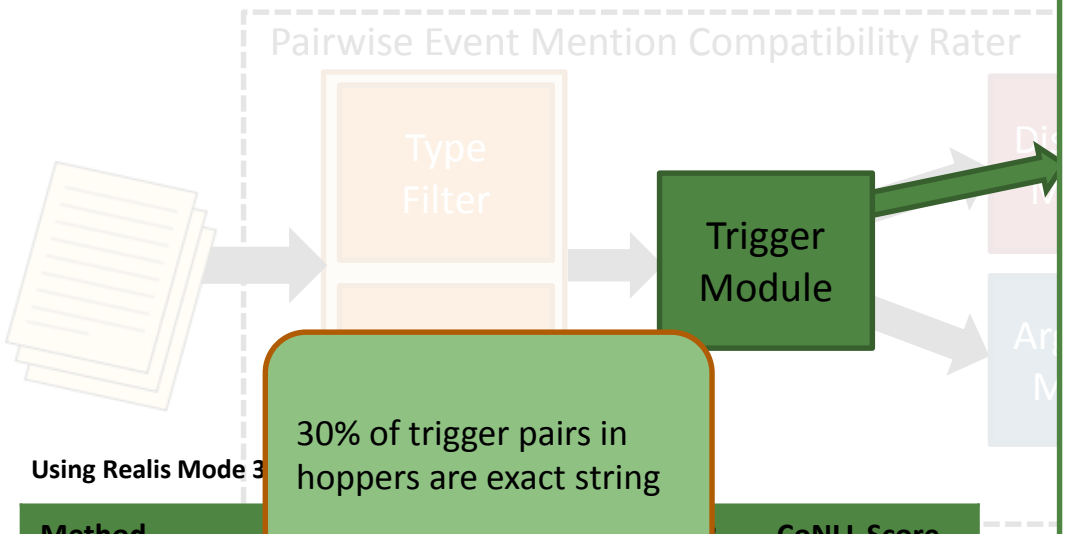
Triggers are compatible

1. …only if they match exactly.
   kills⇔kills **[EXACT]**
2. …if they share a stem.
   indicted ⇔ indicts **[SAME_STEM]**
3. …also if they share a WordNet synset or derived relationship.
   transport ⇔ ship **[SYNONYM]**
   bombings ⇔ bombed
4. …also if they can be linked by a WordNet hypernym relation.
   executed ⇔ hanged **[HYP*NYM]**
5. …also if they are included in a whitelist derived from training.
   death ⇔ fatally **[MANUAL]**
6. …for all pairs of triggers.
   shoot ⇔ impale

Using Realis Mode 3

Method	PairP	PairR	CoNLL Score
All Singletons	0.00	0.00	48.85
T=EXACT	**57.0**	29.1	69.36
T=SAME_STEM	52.4	41.5	72.01
T=SYNONYM	50.2	47.4	72.58
T=HYP*NYM	49.9	49.5	72.13
T=MANUAL	38.0	76.8	**73.44**
All Triggers Compatible	30.4	**88.8**	66.30

Event Hoppers – Faceted Approach



Using Realis Mode 3

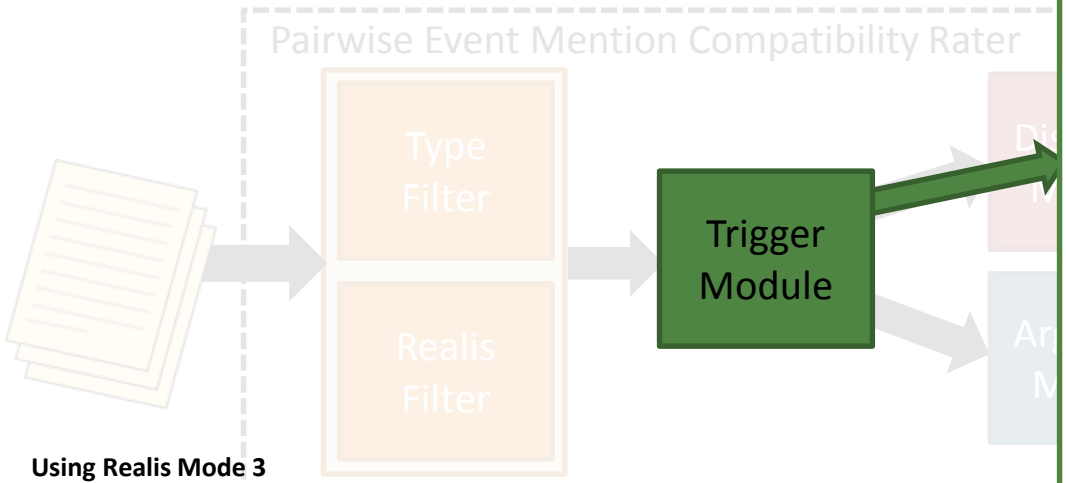
Method			CoNLL Score
All Singletons	0.0	0.00	48.85
T=EXACT	**57.0**	29.1	69.36
T=SAME_STEM	52.4	41.5	72.01
T=SYNONYM	50.2	47.4	72.58
T=HYP*NYM	49.9	49.5	72.13
T=MANUAL	38.0	76.8	**73.44**
All Triggers Compatible	30.4	**88.8**	66.30

Six modes:

Triggers are compatible

1. …only if they match exactly.
   kills⇔kills **[EXACT]**
2. …if they share a stem.
   indicted ⇔ indicts **[SAME_STEM]**
3. …also if they share a WordNet synset or derived relationship.
   transport ⇔ ship **[SYNONYM]**
   bombings ⇔ bombed
4. …also if they can be linked by a WordNet hypernym relation.
   executed ⇔ hanged **[HYP*NYM]**
5. …also if they are included in a whitelist derived from training.
   death ⇔ fatally **[MANUAL]**
6. …for all pairs of triggers.
   shoot ⇔ impale

Event Hoppers – Faceted Approach



Six modes:

Triggers are compatible

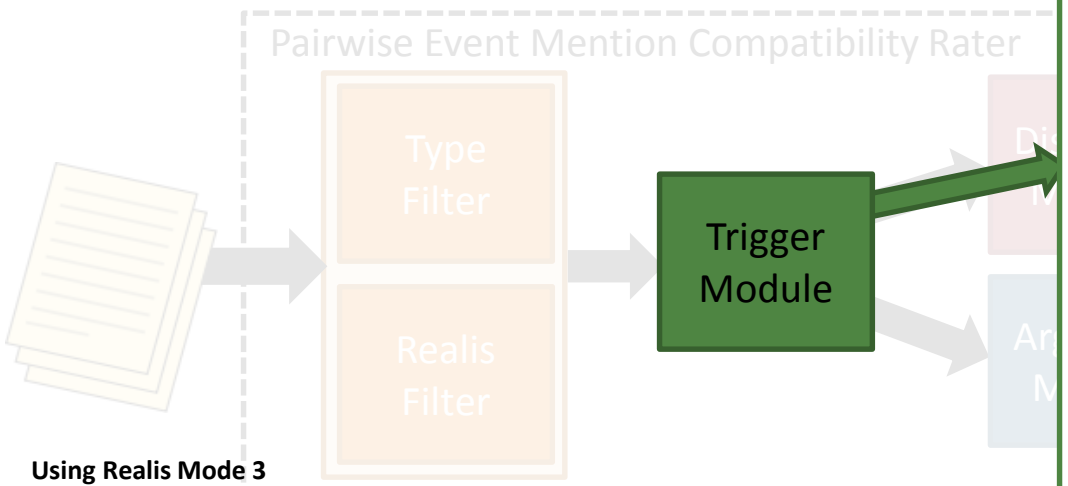
1. …only if they match exactly.
   kills⇔kills **[EXACT]**
2. …if they share a stem.
   indicted ⇔ indicts **[SAME_STEM]**
3. …also if they share a WordNet synset or derived relationship.
   transport ⇔ ship **[SYNONYM]**
   bombings ⇔ bombed
4. …also if they can be linked by a WordNet hypernym relation.
   executed ⇔ hanged **[HYP*NYM]**
5. …also if they are included in a whitelist derived from training.
   death ⇔ fatally **[MANUAL]**
6. …for all pairs of triggers.
   shoot ⇔ impale

Using Realis Mode 3

Method			CoNLL Score
All Singletons			48.85
T=EXACT			69.36
T=SAME_STEM			72.01
T=SYNONYM			72.58
T=HYP*NYM	49.9	49.5	72.13
T=MANUAL	38.0	76.8	**73.44**
All Triggers Compatible	30.4	**88.8**	66.30

Only 50% of triggers in hoppers have a direct relation in WordNet

Event Hoppers – Faceted Approach



Six modes:

Triggers are compatible

1. …only if they match exactly.
   kills⇔kills **[EXACT]**
2. …if they share a stem.
   indicted ⇔ indicts **[SAME_STEM]**
3. …also if they share a WordNet synset or derived relationship.
   transport ⇔ ship **[SYNONYM]**
   bombings ⇔ bombed
4. …also if they can be linked by a WordNet hypernym relation.
   executed ⇔ hanged **[HYP*NYM]**

…also if they are included in a whitelist derived from training.
death ⇔ fatally **[MANUAL]**

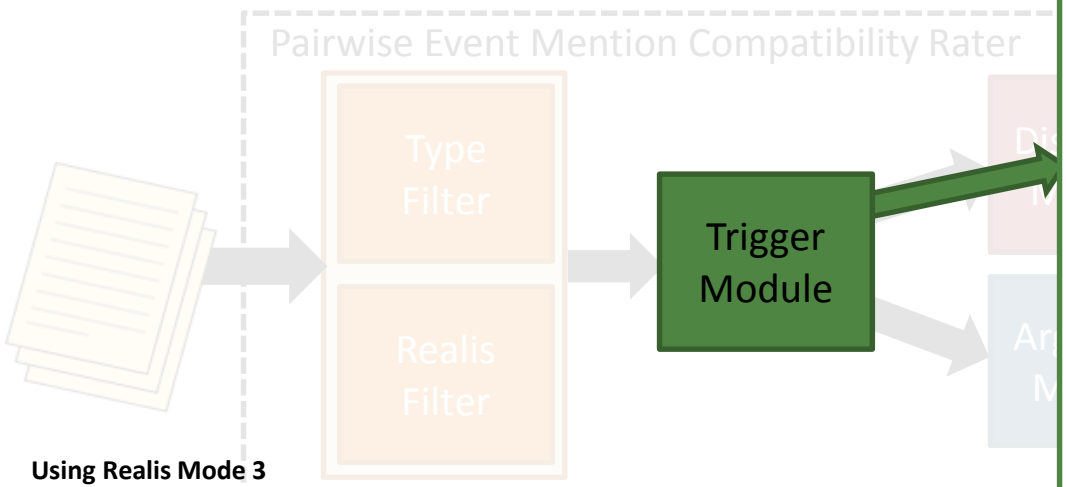
…for all pairs of triggers.
shoot ⇔ impale

Using Realis Mode 3

Method	PairP	PairR	CoNLL Score
All Singletons	0.00	0.00	48.85
T=EXACT	**57.0**	29.1	
T=SAME_STEM	52.4	41.5	
T=SYNONYM	50.2	47.4	
T=HYP*NYM	49.9	49.5	
T=MANUAL	38.0	76.8	**73.44**
All Triggers Compatible	30.4	**88.8**	66.30

Learned lexicon from training data provides good gains

Event Hoppers – Faceted Approach



Six modes:

Triggers are compatible

1. …only if they match exactly.
   kills⇔kills **[EXACT]**
2. …if they share a stem.
   indicted ⇔ indicts **[SAME_STEM]**
3. …also if they share a WordNet synset or derived relationship.
   transport ⇔ ship **[SYNONYM]**
   bombings ⇔ bombed
4. …also if they can be linked by a WordNet hypernym relation.
   executed ⇔ hanged **[HYP*NYM]**
5. …also if they are included in a whitelist derived from training.
   death ⇔ fatally **[MANUAL]**

…for all pairs of triggers.

shoot ⇔ impale

Using Realis Mode 3

Method	PairP	PairR	CoNLL Score
All Singletons	0.00	0.00	48.85
T=EXACT	**57.0**	29.1	69.36
T=SAME_STEM	52.4	41.5	
T=SYNONYM	50.2	47.4	
T=HYP*NYM	49.9	49.5	
T=MANUAL	38.0	76.8	
All Triggers Compatible	30.4	**88.8**	66.30

How can we learn these 12% of triggers are compatible?

Event Hoppers – Faceted Approach

1) Quote linking – for quoted sentences (possibly distant in the document) in forum data [e.g., bolt].
2) Detect chains of terms with same stem.
3) Determine when adjacent pairs in the chain should be linked.
   1) Positive cues – e.g., "the attack" **[POSITIVE]**
   2) Negative cues – e.g., "a different attack" **[POS_NO_NEG]**
   3) Machine learning from cues. **[Discourse ML]**

Discourse Module

Argument Module

Pairwise Selection Module

Event Hoppers

Example Stem-based chain

Amabassador **visits** French researcher in Tehran prison
PARIS, Aug 14, 2009 (AFP)
France's ambassador to Iran on Friday **visited** a young French academic in the Tehran prison where she is being held on spying charges, the foreign ministry said here.
"He explained to her that the French authorities are doing all they can to obtain her release as soon as possible," a spokesman said.
The **visit** was ambassador Bernard Poletti's second trip to Evin prison to see Clotide Reiss, who was among at least 110 defendants tried last week on charges related to huge post-election protests across Iran.

Using Realis Mode 3, Triggers up to Whitelist

Stem-based Chains

Method	PairP	PairR	CoNLL Score
All Singletons	0.00	0.00	37.96
D=POSITIVE	39.9	5.7	43.70
D=POS_NO_NEG	39.9	5.8	43.78
D=ALL	**47.4**	27.9	54.63
Discourse ML	48.9	30.4	54.87
No Discourse	27.1	**58.2**	**55.44**

Task 2

Event Hoppers – Faceted Approach

1) Quote linking – for quoted sentences (possibly distant in the document) in forum data [e.g., bolt].
2) Detect chains of terms with same stem.
3) Determine when adjacent pairs in the chain should be linked.
   1) Positive cues – e.g., "the attack" **[POSITIVE]**
   2) Negative cues – e.g., "a different attack" **[POS_NO_NEG]**
   3) Machine learning from cues. **[Discourse ML]**

Discourse Module

Argument Module

Pairwise Selection Module

Event Hoppers

Example Stem-based chain

Amabassador **visits** French researcher in Tehran prison
PARIS, Aug 14, 2009 (AFP)
France's ambassador to Iran on Friday **visited** a young French academic in the Tehran prison where she is being held on spying charges, the foreign ministry said here.
"He explained to her that the French authorities are doing all they can to obtain her release as soon as possible," a spokesman said.
The **visit** was ambassador Bernard Poletti's second trip to Evin prison to see Clotide Reiss, who was among at least 110 defendants tried last week on charges related to huge post-election protests across Iran.

Using Realis Mode 3, Triggers up to Whitelist

Stem-based Chains

Method	PairP	PairR	CoNLL Score
All Singletons	0.00	0.00	48.85
D=POSITIVE	50.7	8.6	57.05
D=POS_NO_NEG	50.9	8.5	56.94
D=ALL	53.6	31.3	68.93
Discourse ML	**59.5**	35.4	70.05
No discourse	38.0	**76.8**	**73.44**

Task 3

Event Hoppers – Faceted Approach

1) Quote linking – for quoted sentences (possibly distant in the document) in forum data [e.g., bolt].
2) Detect chains of terms with same stem.
3) Determine when adjacent pairs in the chain should be linked.
   1) Positive cues – e.g., "the attack" **[POSITIVE]**
   2) Negative cues – e.g., "a different attack" **[POS_NO_NEG]**
   3) Machine learning from cues. **[Discourse ML]**

Discourse Module

Argument Module

Pairwise Selection Module

Event Hoppers

Example Stem-based chain

Amabassador **visits** French researcher in Tehran prison
PARIS, Aug 14, 2009 (AFP)
France's ambassador to Iran on Friday **visited** a young French academic in the Tehran prison where she is being held on spying charges, the foreign ministry said here.
"He explained to her that the French authorities are doing all they can to obtain her release as soon as possible," a spokesman said.
The **visit** was ambassador Bernard Poletti's second trip to Evin prison to see Clotide Reiss, who was among at least 110 defendants tried last week on charges related to huge post-election protests across Iran.

Only 9% of pairs have explicit discourse cue, and negative cues are minimal

Using Realis Mode 3
Triggers up to Whiteli

Method			CoNLL Score
All Singletons	0.00	0.00	48.85
D=POSITIVE	50.7	8.6	57.05
D=POS_NO_NEG	50.9	8.5	56.94
D=ALL	53.6	31.3	68.93
Discourse ML	**59.5**	35.4	70.05
No discourse	38.0	**76.8**	**73.44**

Task 3

Event Hoppers – Faceted Approach

1) Quote linking – for quoted sentences (possibly distant in the document) in forum data [e.g., bolt].
2) Detect chains of terms with same stem.
3) Determine when adjacent pairs in the chain should be linked.
   1) Positive cues – e.g., "the attack" **[POSITIVE]**
   2) Negative cues – e.g., "a different attack" **[POS_NO_NEG]**
   3) Machine learning from cues. **[Discourse ML]**

Discourse Module

Argument Module

Pairwise Selection Module

Event Hoppers

Example Stem-based chain

Amabassador **visits** French researcher in Tehran prison
PARIS, Aug 14, 2009 (AFP)
France's ambassador to Iran on Friday **visited** a young French academic in the Tehran prison where she is being held on spying charges, the foreign ministry said here.
"He explained to her that the French authorities are doing all they can to obtain her release as soon as possible," a spokesman said.
The **visit** was ambassador Bernard Poletti's second trip to Evin prison to see Clotide Reiss, who was among at least 110 defendants tried last week on charges related to huge post-election protests across Iran.

Using Realis Mode 3, Triggers up to Whitelist

Stem-based Chains

Method	PairP	PairR	CoNLL Score
All Singletons	0.00	0.00	48.8
D=POSITIVE	50.7	8.6	57.0
D=POS_NO_NEG	50.9	8.5	56.9
D=ALL	53.6	31.3	68.95
Discourse ML	**59.5**	35.4	70.05
No discourse	38.0	**76.8**	**73.44**

Improving hopperation with discourse model is an open research question

Task 3

Event Hoppers – Faceted Approach

Temporal Arg Matching

1) Normalize Relative Times
2) Calculate Start/End points
3) Detect overlap of spans

"last week" ——————
"last Tuesday" ——

Spatial Arg Matching

1) Link into gazetteer
2) If both can be linked, search for containment relation.

General Arg Matching

1) Extract arguments using in-house SRL.
2) Convert to named roles (e.g., "victim", "attacker") if possible
3) Detect compatibility between args with same role – strict, moderate, or weak.

Strict: Exact match, Entity Coref (heads), Same number, Same WordNet synset (after WSD)
Moderate: Partial string match, Same WordNet synset (no WSD), WordNet hypernyms (after WSD), Mismatched number, Compatible entity types (nominal)
Weak: One has number, Entity Coref (any), WordNet hypernyms (no WSD)

Argument Module

Filter

Pairwise Selection Module

Event Hoppers

Using Realis Mode 3, Triggers up to Whitelist

Only 18% of triggers have any argument match, and the precision is 54%

Method				
All Singletons				
Require Strict Arg Match	**[REQ_HIGH]**			
Require Moderate Arg Match	**[REQ_MED]**	56.9	17.1	63.27
Require Weak Arg Match	**[REQ_LOW]**	54.4	18.0	63.20
Prohibit Any Mismatch	**[NO_MISS]**	47.3	54.9	73.18
Prohibit Multiple Mismatch	**[NO_MULTI]**	38.2	73.7	73.33
Prohibit Spatio-Temporal Mismatch	**[SPACETIME]**	38.5	73.1	73.25
Accept All		38.0	**76.8**	**73.44**

Event Hoppers – Faceted Approach

Temporal Arg Matching

1) Normalize Relative Times
2) Calculate Start/End points
3) Detect overlap of spans

"last week" ——————
"last Tuesday" ——

Spatial Arg Matching

1) Link into gazetteer
2) If both can be linked, search for containment relation.

General Arg Matching

1) Extract arguments using in-house SRL.
2) Convert to named roles (e.g., "victim", "attacker") if possible
3) Detect compatibility between args with same role – strict, moderate, or weak.

Strict: Exact match, Entity Coref (heads), Same number, Same WordNet synset (after WSD)
Moderate: Partial string match, Same WordNet synset (no WSD), WordNet hypernyms (after WSD), Mismatched number, Compatible entity types (nominal)
Weak: One has number, Entity Coref (any), WordNet hypernyms (no WSD)

Argument Module

Pairwise Selection Module

Event Hoppers

Using Realis Mode 3, Triggers up to Whitelist

Task 3

Prohibiting mismatches helps P, hurts R, same F

Method				
All Singletons				
Require Strict Arg Match	**[REQ_HIGH]**			
Require Moderate Arg Match	**[REQ_MED]**			
Require Weak Arg Match	**[REQ_LOW]**	54.4	18.0	63.20
Prohibit Any Mismatch	**[NO_MISS]**	47.3	54.9	73.18
Prohibit Multiple Mismatch	**[NO_MULTI]**	38.2	73.7	73.33
Prohibit Spatio-Temporal Mismatch	**[SPACETIME]**	38.5	73.1	73.25
Accept All		38.0	**76.8**	**73.44**

Event Hoppers – Faceted Approach

Temporal Arg Matching

1) Normalize Relative Times
2) Calculate Start/End points
3) Detect overlap of spans

"last week"

"last Tuesday"

Spatial Arg Matching

1) Link into gazetteer
2) If both can be linked, search for containment relation.

General Arg Matching

1) Extract arguments using in-house SRL.
2) Convert to named roles (e.g., "victim", "attacker") if possible
3) Detect compatibility between args with same role – strict, moderate, or weak.

Strict: Exact match, Entity Coref (heads), Same number, Same WordNet synset (after WSD)
Moderate: Partial string match, Same WordNet synset (no WSD), WordNet hypernyms (after WSD), Mismatched number, Compatible entity types (nominal)
Weak: One has number, Entity Coref (any), WordNet hypernyms (no WSD)

Argument Module

Pairwise Selection Module

Event Hoppers

Using Realis Mode 3, Triggers up to Whitelist

Task 3

Method		PairP		
All Singletons		0.00		
Require Strict Arg Match	**[REQ_HIGH]**	**68.1**		
Require Moderate Arg Match	**[REQ_MED]**	56.9		
Require Weak Arg Match	**[REQ_LOW]**	54.4	18.0	.20
Prohibit Any Mismatch	**[NO_MISS]**	47.3	54.9	73.18
Prohibit Multiple Mismatch	**[NO_MULTI]**	38.2	73.7	73.33
Prohibit Spatio-Temporal Mismatch	**[SPACETIME]**	38.5	73.1	73.25
Accept All		38.0	**76.8**	**73.44**

36% of triggers with no matches or mismatches

Event Hoppers – Faceted Approach

Tiered Trigger/Argument Model

Exact Match Same Stem Synonym/Derived	**Prohibit Multiple Mismatch**
Whitelisted	**Machine Learning**
Other	**Require Strict Arg Match**

Machine Learning Model

Separate Models for StemMatched and nonStemMatched

Features: Trigger Agreement Type, Lexical Pairs, Realis Pairs, Typed Argument Matches, Argument Existence

Argument Module

Pairwise Selection Module

Event Hoppers

Using Realis Mode 3,
Triggers up to Whitelist (for non-Tiered)

Task 2

Method	PairP	PairR	CoNLL Score
All Singletons	0.00	0.00	37.96
Tiered Model	32.3	44.2	55.54
Tiered Model + Discourse ML	35.4	36.1	53.50
Argument ML	38.6	36.1	55.17
Argument ML + Discourse ML	45.8	31.8	55.22
Accept All Triggers/Pairs	27.1	58.2	55.44

Different Models perform equally well for Task 2

Event Hoppers – Faceted Approach

Tiered Trigger/Argument Model

Exact Match Same Stem Synonym/Derived	**Prohibit Multiple Mismatch**
Whitelisted	**Machine Learning**
Other	**Require Strict Arg Match**

Machine Learning Model

Separate Models for StemMatched and nonStemMatched

Features: Trigger Agreement Type, Lexical Pairs, Realis Pairs, Typed Argument Matches, Argument Existence

Argument Module

Pairwise Selection Module

Event Hoppers

Using Realis Mode 3,
Triggers up to Whitelist (for non-Tiered)

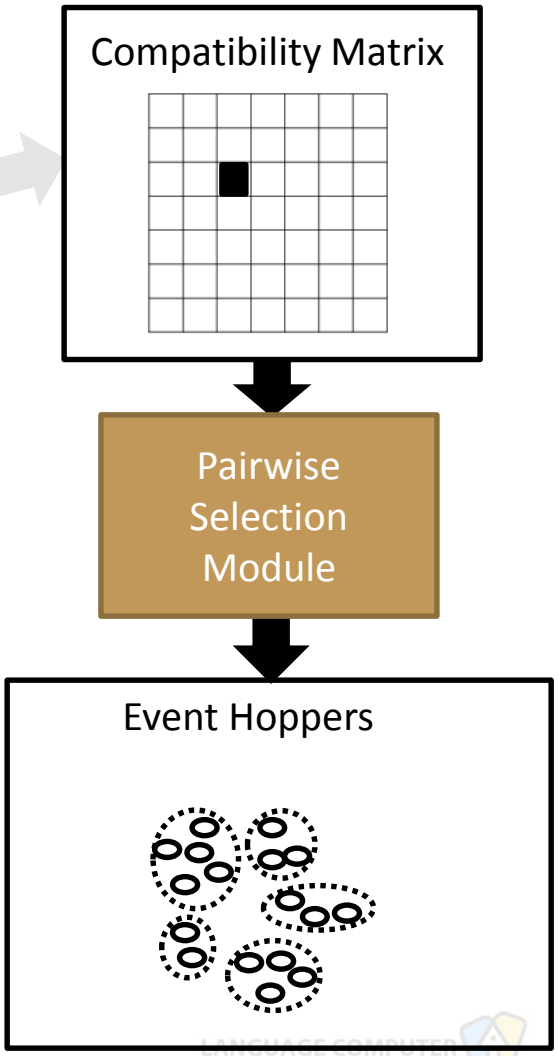
Task 3

Method	PairP	PairR	CoNLL Score
All Singletons	0.00	0.00	48.85
Tiered Model	44.4	53.8	71.78
Tiered Model + Discourse ML	48.1	47.1	69.96
Argument ML	50.3	39.1	70.42
Argument ML + Discourse ML	54.4	36.3	70.20
Accept All Triggers/Pairs	38.0	76.8	73.44

Argument and Discourse Models don't help for Task 3

Event Hoppers – Faceted Approach

1) Results of Type, Realis, Trigger, Discourse, and Argument Components converted into event-event compatibility scores
   a) Incompatibilities are treated as infinitely negative
   b) Discourse-based compatibility is heavily weighted.
   c) Argument compatibilities are additive (more argument overlap increases the evidence for event compatibility).
2) Each event starts in its own hopper.
3) Greedily find the hoppers associated with the highest scoring pair of events (positive scores only).
4) If there are no known incompatibilities between any pair of events within these two hoppers, merge them into one hopper.
5) Stop when everything is merged or incompatible.



Event Hoppers – Results

Task 2

Methods (Representative Selection, Ordered by decreasing recall)	PairP	PairR	CoNLL Score
All Singletons (Baseline)	0.00	0.00	37.96
All Events (Baseline)	15.6	65.6	46.65
R=STRICT	23.6	63.3	50.69
R=STRICT, T=MANUAL	27.1	58.2	55.44
Tiered Model, R=GENERIC, D=POSITIVE **(Task 2: Run 2)**	**30.7**	**45.2**	**54.98**
Tiered Model: No Discourse, R=STRICT	32.3	44.2	55.54
R=STRICT, T=MANUAL, A=NO_MISS	30.7	42.4	55.89
R=STRICT, T=SYNONYM	35.3	38.2	56.59
ML Model: No Discourse, R=STRICT, T=MANUAL	38.6	36.1	55.17
R=GENERIC, T=SYNONYM, D=POS NO NEG, A=SPACE TIME **(Task 2: Run 1,3)**	**28.2**	**35.7**	**56.54**
R=STRICT, T=SAME_STEM	35.6	34.7	55.69
R=STRICT, T=MANUAL, D=ALL (Stem-based Chains)	47.4	27.9	54.63
R=STRICT, T=EXACT	39.6	27.0	54.72
R=STRICT, T=MANUAL, A=REQ_LOW	51.4	14.5	50.69

Event Hoppers – Results

Task 3

Methods (Representative Selection, Ordered by decreasing recall)	PairP	PairR	CoNLL Score
All Singletons (Baseline)	0.00	0.00	48.85
All Events (Baseline)	18.3	99.1	57.52
R=STRICT	30.4	88.8	66.30
R=STRICT, T=MANUAL [High Recall] (Task 3: Run 3)	**38.0**	**76.8**	**73.44**
R=STRICT, T=MANUAL, A=NO_MISS	47.3	54.9	73.19
Tiered R=STRICT, D:ALL, A:TIERED [Balanced Precision/Recall] (Task 3: Run 2)	**49.0**	**54.1**	**72.84**
Tiered Model: No Discourse, R=STRICT	44.4	53.8	71.78
R=STRICT, T=SYNONYM	50.2	47.4	72.58
R=STRICT, T=SAME_STEM	52.4	41.5	72.01
ML Model: No Discourse, R=STRICT, T=MANUAL	50.3	39.1	70.42
Arg ML + Discourse ML, R=STRICT, T=MANUAL [High Precision] (Task 3: Run 1)	**51.5**	**38.8**	**70.87**
R=STRICT, T=MANUAL, D=ALL (Stem-based Chains)	53.6	31.3	68.93
R=STRICT, T=EXACT	57.0	29.1	69.36
R=STRICT, T=MANUAL, A=REQ_LOW	54.4	18.0	63.20

LANGUAGE COMPUTER

Event Hoppers – Evaluation Results

Task 2

Methods (Representative Selection, Ordered by decreasing recall)	CoNLL Score	(Test)
Run 1 – R=GENERIC, T=SYNONYM, D=POS NO NEG, A=SPACE TIME	62.80	56.54
Run 2 – Tiered Model, R=GENERIC, D=POSITIVE	62.95	54.98
Run 3 – R=GENERIC, T=SYNONYM, D=POS NO NEG, A=SPACE TIME	62.63	

Task 3

Methods (Representative Selection, Ordered by decreasing recall)	CoNLL Score	(Test)
Run 1 – Argument ML + Discourse ML, R=STRICT, T=MANUAL	71.86	70.87
Run 2 – Tiered Model, R=STRICT, D:ALL, A:TIERED	74.87	72.84
Run 3 – R:STRICT, T:MANUAL	75.69	73.44

Event Hoppers – Conclusions

1. Realis has a significant impact in improving precision.
2. Argument matching was shown to be difficult to incorporate properly
   a. Requiring an argument to match significantly drops recall – many events have no arguments OR have arguments which could not be extracted properly.
   b. Prohibiting mismatched arguments does not impact the score significantly. More attention needs to be paid to this issue.
3. Discourse-based modeling has been shown to perform well stand-alone, but not significantly improve results over high-recall, trigger-based approaches.
4. Scoring bias is towards high recall – better to over-merge than under-merge.
5. Spatio-temporal cues (especially conflicting or compatible ones) were rare.

Conclusions

- Found core of strategies which work well for both tasks
  - More research to incorporate the other pieces
- Demo
  - LCC's KB populated with the event nugget data and hoppers
- Questions?