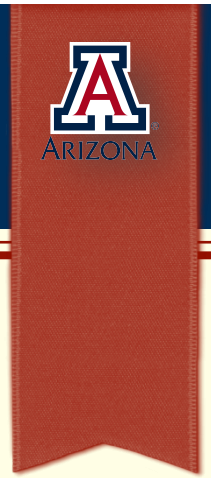


IDENTIFYING LATENT ATTRIBUTES FROM VIDEO SCENES USING KNOWLEDGE ACQUIRED FROM LARGE COLLECTIONS OF TEXT DOCUMENTS

Anh Xuan Tran
April 30, 2014

Dissertation Committee:
Paul Cohen, Mihai Surdeanu, Kobus Barnard, Ken McAllister

Outline

Defining the Problem

Models

Performance Measures

Results

Conclusions / Future Work

Defining the Problem

Models

Performance Measures

Results

Conclusions / Future Work

Motivation

- **Imagine two young children chasing each other in the playground on a bright sunny day.**
- Are they enjoying themselves?
- Is the child running for his life or is he happy and playful?
- It's a chase! Should someone notify the police?

- How were you able to answer these questions?
  - Needed latent (or hidden) attributes about the scene.

Motivation: Surveillance Application

- Should someone notify the police?

- A surveillance system could automatically dispatch the police if it can infer that someone is in a state of distress.

Problem Definition

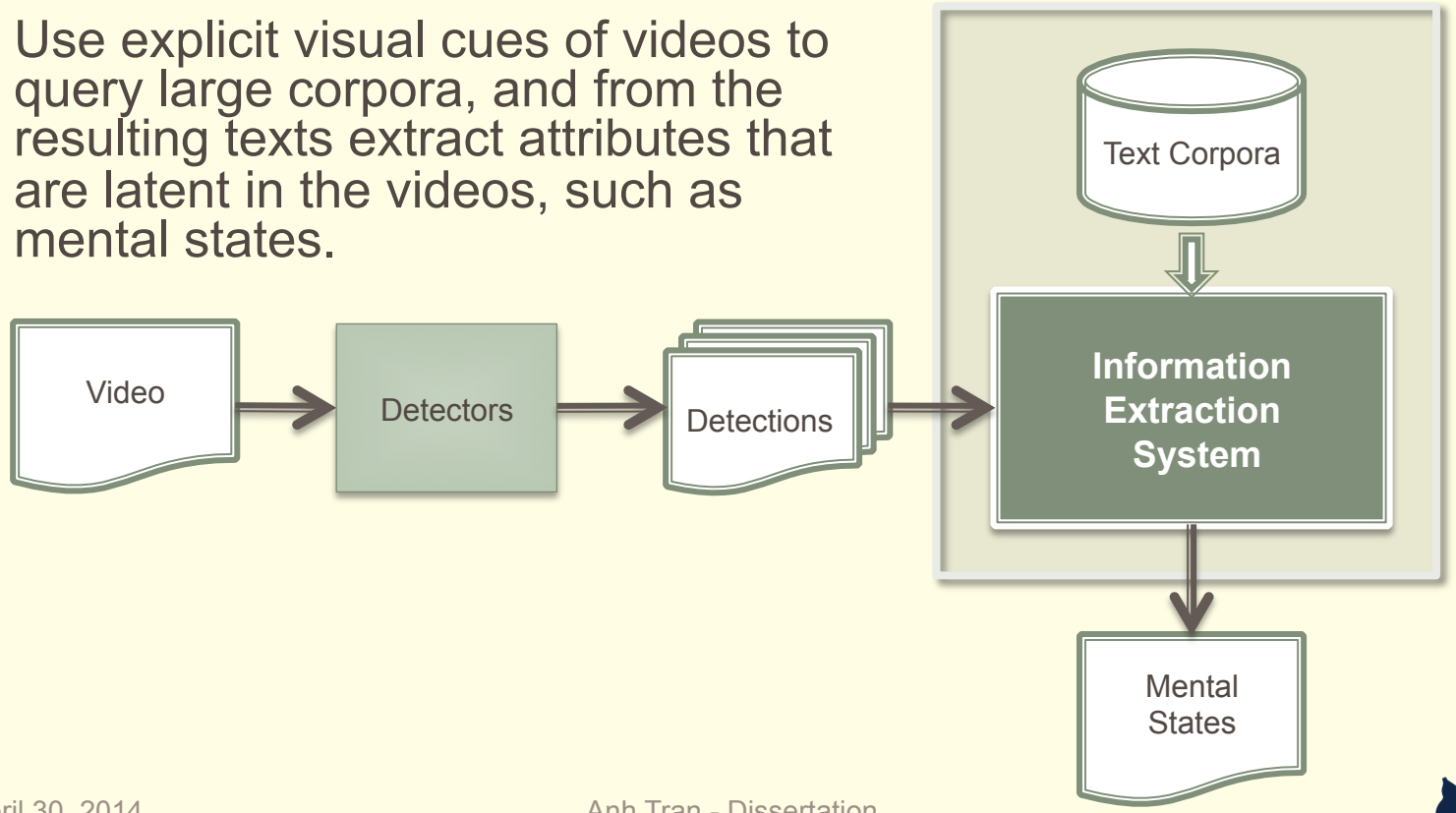
Identifying latent attributes from video scenes, with a focus on the mental states of activity participants, given some contextual information about the scenes.

- **Latent attributes** are elements describing a scene that cannot be directly observed or extracted from the scene.
- Automatic identification of latent attributes is a challenging task.
  - Machines do not have access to the same wisdom that humans do.
  - It only knows what is told, e.g., that there is a chase and it might involve a child.
  - So how to identify latent information?

The Approach

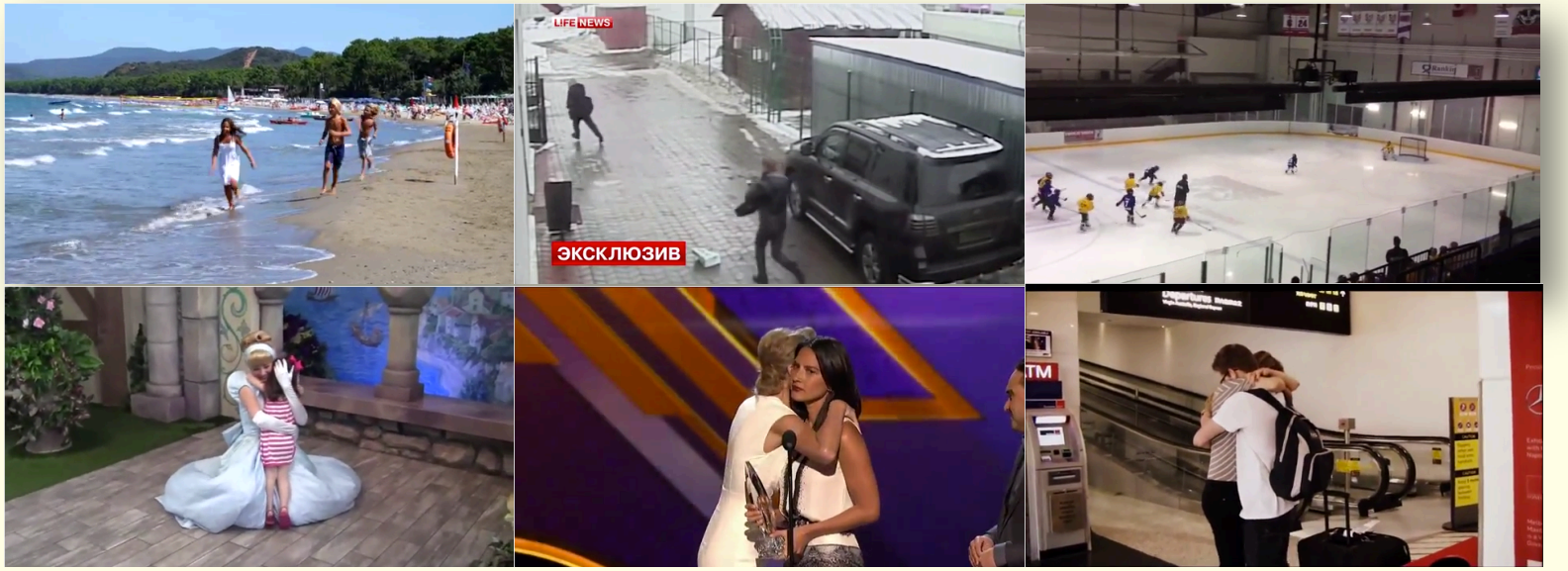
Attributes that are latent in visual representation are often explicit in textual representation.

- Use explicit visual cues of videos to query large corpora, and from the resulting texts extract attributes that are latent in the videos, such as mental states.



Video Datasets

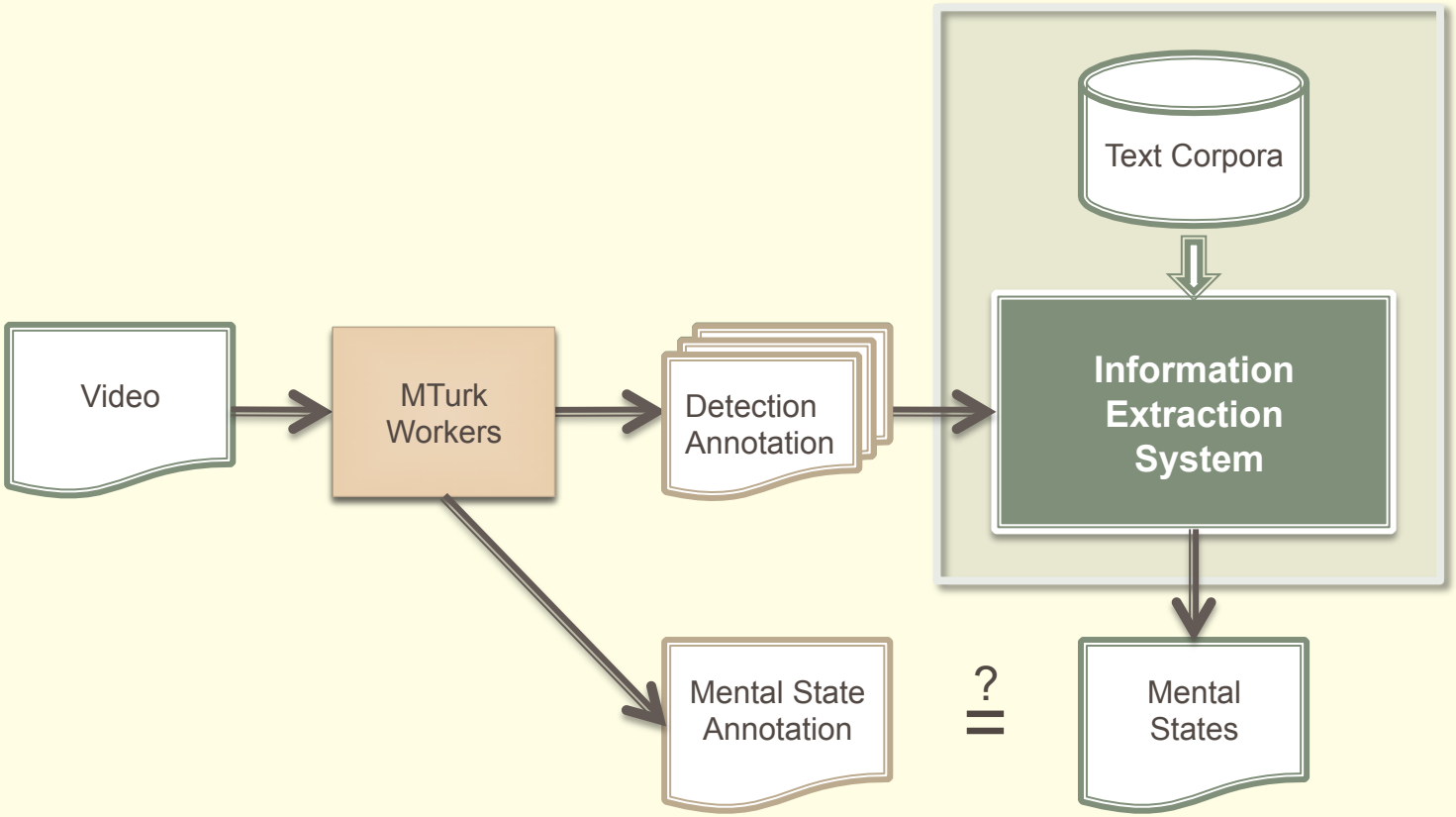
- Generated 2 datasets: *chase* (26) and *hug* (45)



Video Annotations

- Used MTurk to generate ground-truth annotations
- Annotate mental states of actors → Use as gold-standard set for evaluation
  - Ask 10 workers, aggregate results to form frequency distribution
- Annotate visual cues – humans as proxy for human-level detection system
  - Select from predefined list of tags (e.g., policeman, children, walk, run, etc.)
  - Input into system as contextual knowledge about the scene

Video Annotations



Text Corpora

- English Gigaword 5th edition corpus [primary]
  - Comprehensive archive of **newswire** articles
  - 26 GB in size
  - Contains 9,876,086 documents, over 4B words
- Google Web N-gram corpus (1-T)
  - *n*-grams up to 5-grams
  - Generated from 1 trillion word tokens from web pages
- Twitter
  - Collected 200 MB streaming Tweets for 1 month
  - Mostly irrelevant Tweets, e.g., “why do I keep chasing after bad boys #smh”
- World Wide Web [future work]
  - Hard to access live web – search APIs provide limited queries
  - ClueWeb09 – 1 billion pages from 2009 – too large for indexing at moment

Defining the Problem

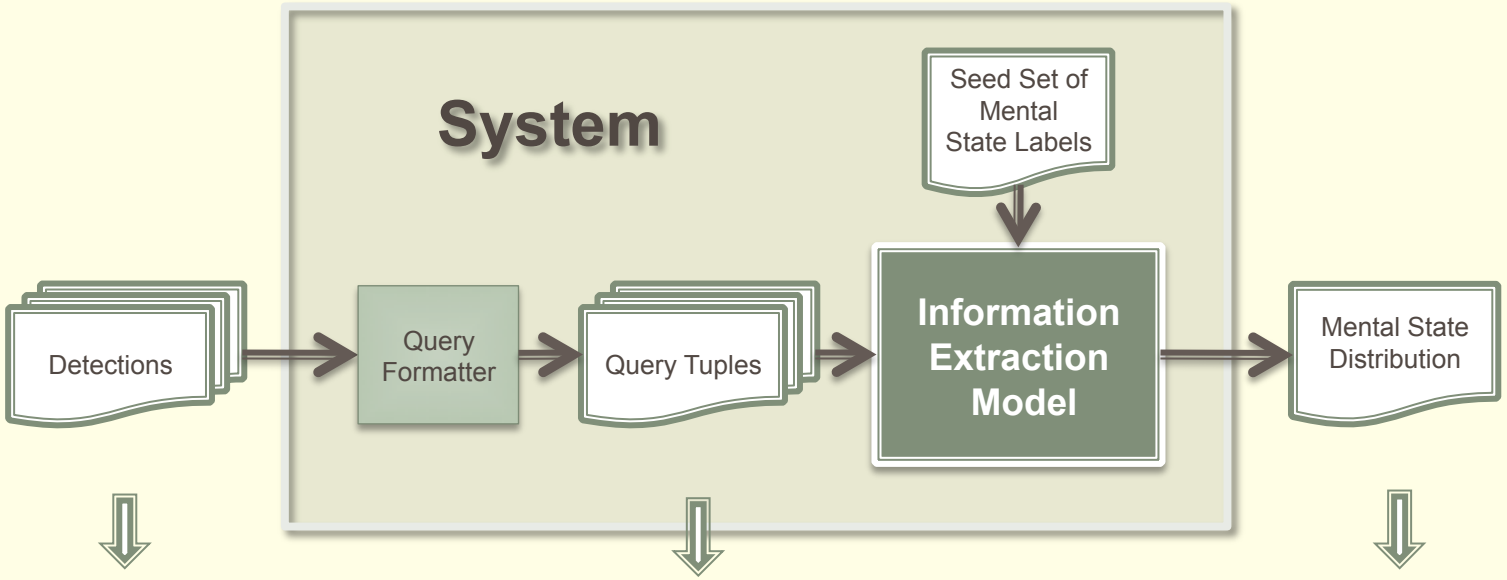
Models

Performance Measures

Results

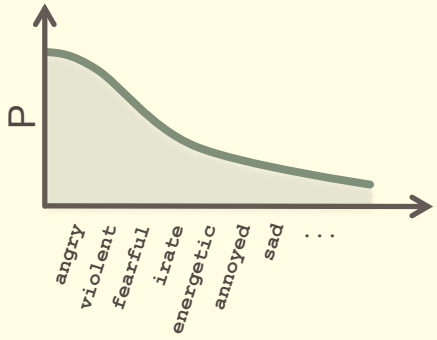
Conclusions / Future Work

System Overview




```
chase
run
...
child
policeman
person
...
```

```
(chase, policeman)
(chase, officer)
(chase, child)
...
```



Seed Set of Mental State Labels

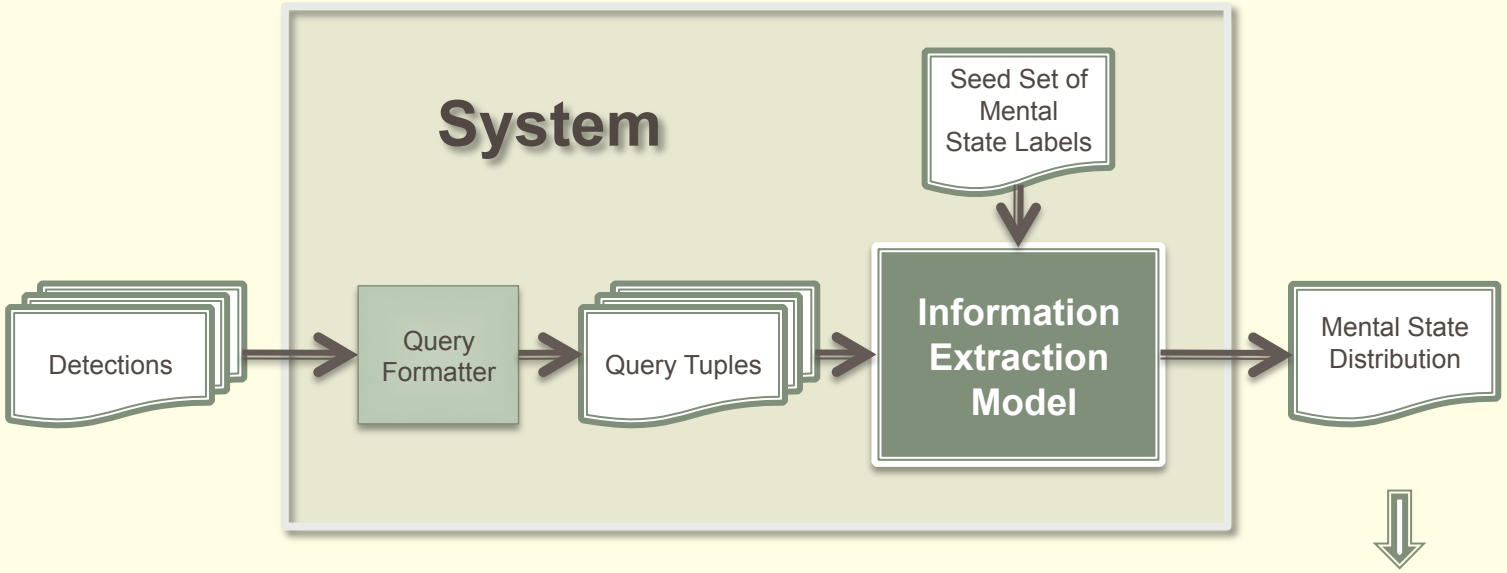
- Contains 160 mental state labels
- Generated from popular dictionaries
  - Profile of Mood States (POMS) psychological test
  - Plutchik's wheel of emotions
  - Others

Source	Example Mental State Labels
POMS	alert, annoyed, energetic, exhausted, helpful, sad, terrified, unworthy, weary, etc.
Plutchik	angry, disgusted, fearful, joyful/joyous, sad, surprised, trusting, etc.
Others	agitated, competitive, cynical, disappointed, excited, giddy, happy, inebriated, violent, etc.

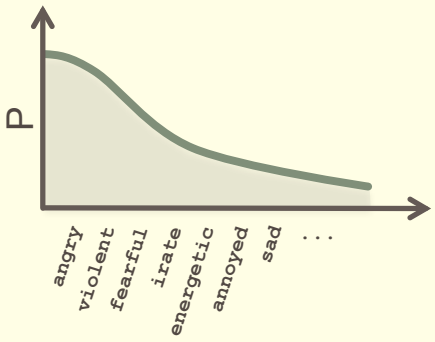
Actor Types

- Limited set of different actor-type detections:
- Inspired by what could reasonably be expected from current state-of-the-art in computer vision. For example:
  - *policeman*: An object detector (Felzenszwalb et al. 2008) can be trained to detect the distinctive uniforms of police officers.
  - *child*: A child can be distinguished by their height, which can be provided under a 3D tracking model (Brau et al. 2013).
- Use synonyms in WordNet to expand set of detections.
  - policeman → policeman, officer

Neighborhood Models

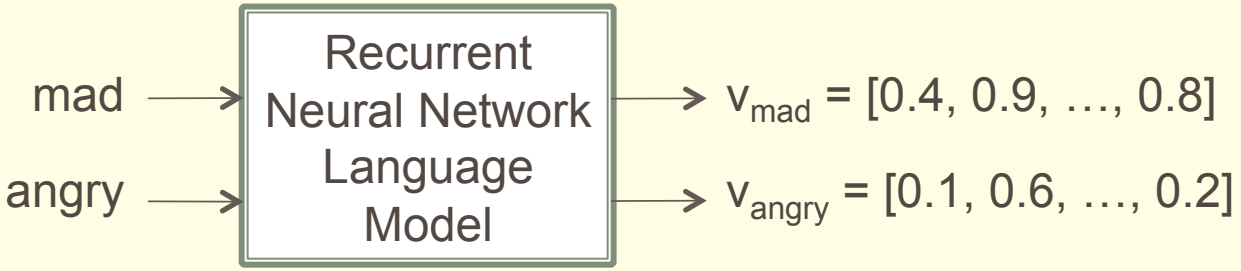


Three largely unsupervised information extraction models.

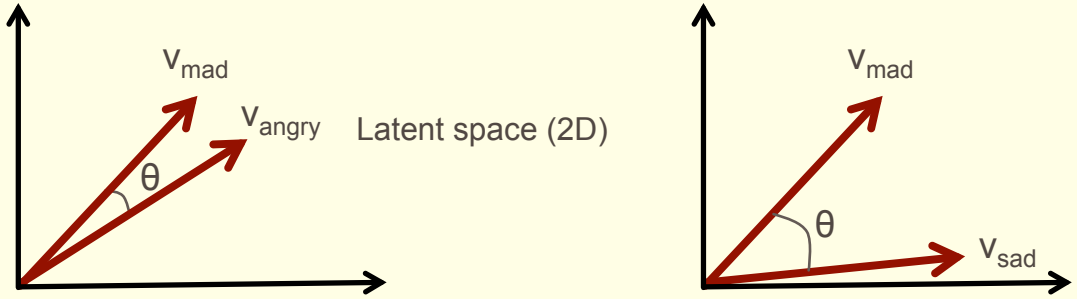


Vector Space with Back-off Linear Interpolation

- **Idea:** Project mental state labels and search context into latent conceptual space produced by a RNNLM (Mikolov et al., 2013a).

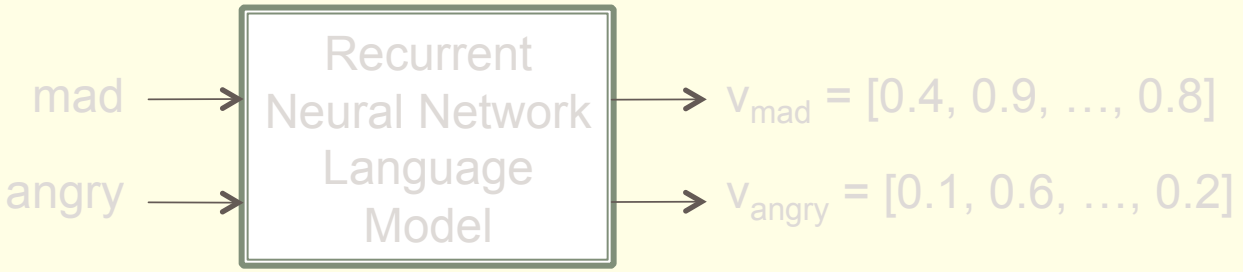


- Compare in latent space using the angle between the vectors.

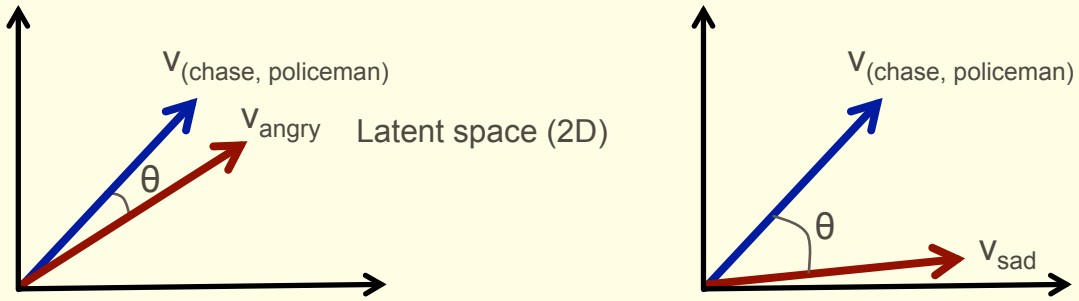


Vector Space with Back-off Linear Interpolation

- **Idea:** Project mental state labels and search context into latent conceptual space produced by a RNNLM (Mikolov et al., 2013a).



- Compare mental state labels to query tuple in latent space.



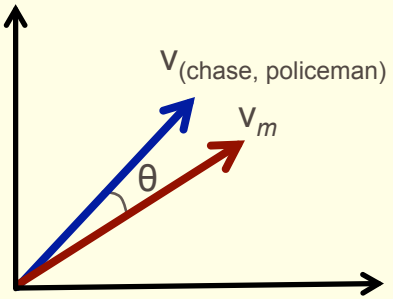
Vector Space with Back-off Linear Interpolation

- Compute context-vector for query tuple:

$$vec(chase, policeman) = vec(chase) + vec(policeman)$$

- Compute similarity to each mental state *m*:

$$cos(\Theta_m) = \frac{vec(m) \cdot vec(context\ tuple)}{||vec(m)||\ ||vec(context\ tuple)||}$$



- → 160 scores per context (or query) tuple.
- Normalize scores to generate a distribution per tuple, average across tuples to create one distribution, and prune to yield final response distribution.

Sentence Co-occurrence with Deleted Interpolation

- **Idea**: Words in the same sentence are likely to be related.
- Rank mental state labels based on the likelihood that they appear in sentences cued by query tuples.
- Interested in the conditional probability:

$$P(m|activity, actor\text{-}type) = \frac{f(m, activity, actor\text{-}type)}{f(activity, actor\text{-}type)}$$

- Normally, we could compute this probability based on relative frequencies.
- However, estimation is unreliable due to sparse data!

Sentence Co-occurrence with Deleted Interpolation

- Cannot estimate probability of trigrams reliably from the corpus, so we estimate probability as linear interpolation of unigrams, bigrams, trigrams.
- Define maximum likelihood probabilities $\hat{P}$ based on relative frequencies:

$$\text{Unigram:} \quad \hat{P}(m) = \frac{f(m)}{N}$$

$$\text{Bigram:} \quad \hat{P}(m|activity) = \frac{f(m, activity)}{f(activity)}$$

$$\text{Trigram:} \quad \hat{P}(m|activity, actor\text{-}type) = \frac{f(m, activity, actor\text{-}type)}{f(activity, actor\text{-}type)}$$

  - N = total number of tokens in the corpus
  - *f*(*m*, *activity*) = number of sentences containing both *m* as an adjective and *activity* as a verb

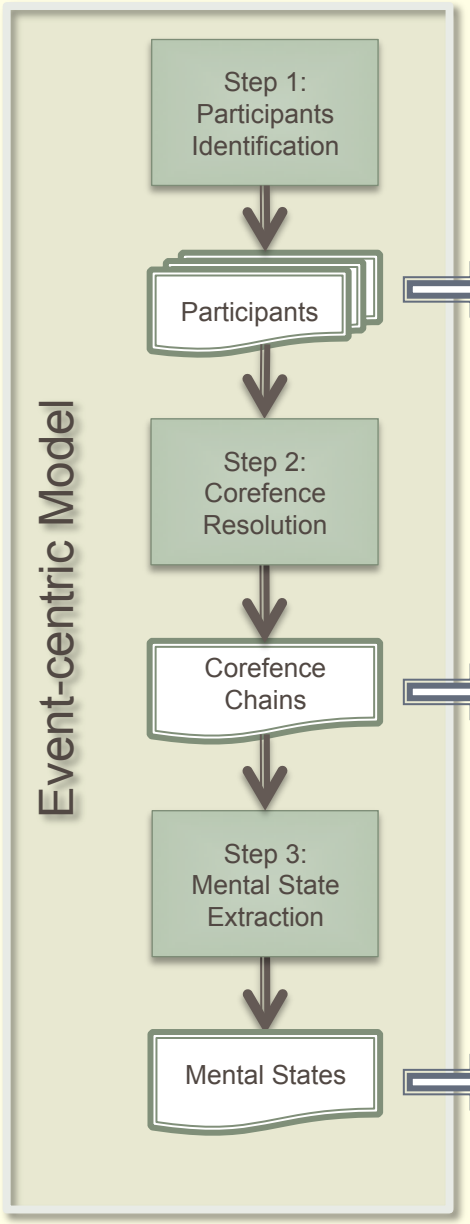
$$P(m|activity, actor\text{-}type) = \lambda_1\hat{P}(m) + \lambda_2\hat{P}(m|activity) + \lambda_3\hat{P}(m|activity, actor\text{-}type)$$

- Use deleted interpolation to estimate lambdas.
- 160 trigram probabilities for each query tuple, average across all query tuples and prune to yield final response distribution.

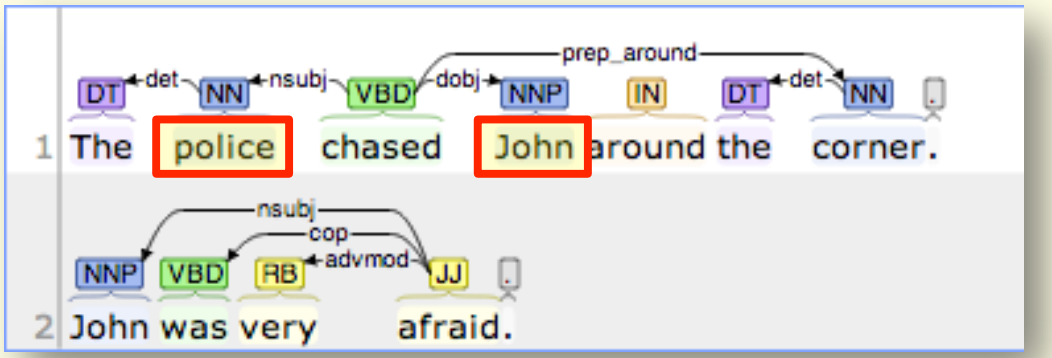
Event-centric with Deleted Interpolation

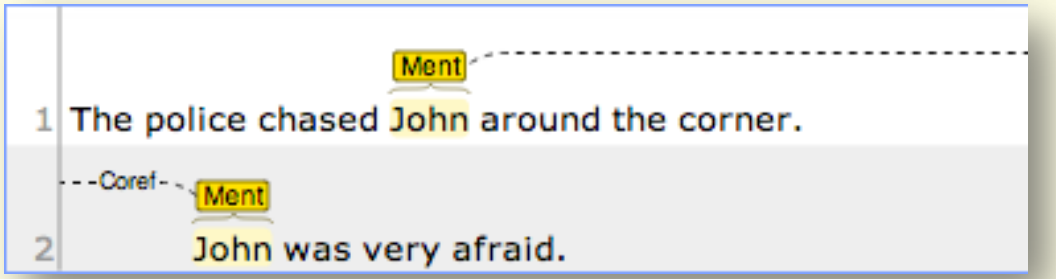
- **Idea:** Identify the event + its participants in the relevant sentences and focus only on the mental states of event participants.

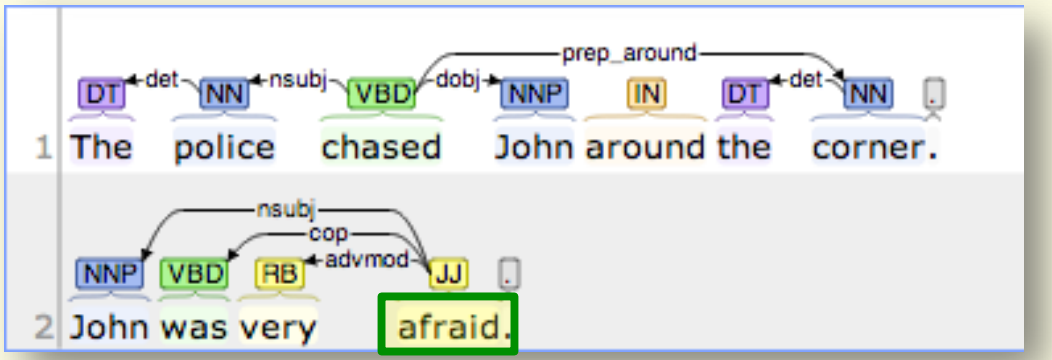
- A smarter, more robust, way to find collocating mental states for joint frequency estimation.
  - Go beyond sentence boundary.
  - Focus on mental states of participants.



Example Output







Defining the Problem

Models

Performance Measures

Results

Conclusions / Future Work

Evaluation

- New task → No standard performance measure
- Compare two normalized distributions over mental states.
- Similarity of **distribution shapes**
  - A good measure must account for the similarity between the shapes of the two distributions (i.e., ratios between weights)
- Semantic similarity of **distribution elements**
  - A good measure must allow for semantic comparisons at the level of distribution elements (i.e., recognize that irate and angry are similar)

Gold G	$(\text{angry}, 0.9), (\text{afraid}, 0.05), (\text{guilty}, 0.05)$
Response R_1	$(\text{angry}, 0.1), (\text{afraid}, 0.2), (\text{guilty}, 0.7)$
Response R_2	$(\text{irate}, 0.45), (\text{mad}, 0.45), (\text{scared}, 0.05), (\text{guilty}, 0.05)$

Known Distribution Similarity Measures

- Many known methods for comparing distributions (Rubner 2000)
- Kullback-Leibler divergence

$$D_{KL}(G||R) = \sum_i G(w_i) \log \frac{G(w_i)}{R(w_i)}$$

- Jeffrey divergence

$$D_J(G||R) = \sum_i G(w_i) \log \frac{G(w_i)}{m_i} + R(w_i) \log \frac{R(w_i)}{m_i}, \quad m_i = \frac{G(w_i) + R(w_i)}{2}$$

- χ^2 statistics

$$D_{\chi^2}(G, R) = \sum_i \frac{(G(w_i) - m_i)^2}{m_i}$$

- Earth Mover's Distance (EMD)
  - Given two collections of dirt piles, EMD measures the least amount of work needed to rearrange the dirt piles of one collection to match the other.
  - Provide different costs for moving a unit of dirt between different piles.
  - Finds optimal solution in super-cubic time.

Constrained Weighted Similarity-Aligned F1

- We start with the standard F_1 measure.

$$precision = \frac{|R \cap G|}{|R|}\ ,\ recall = \frac{|R \cap G|}{|G|}\ ,\ F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \qquad F_1$$

- Generalize the formulas to address our criteria.

$$precision = \frac{1}{|R|} \sum_{r \in R} \max_{g \in G} \sigma(r, g) \qquad \sigma(r, g) = \begin{cases} 1\ , & \text{if } r = g \\ 0\ , & \text{otherwise} \end{cases} \qquad \text{SA-F}_1$$

$$= \sum_{r \in R} R(r) \cdot \max_{g \in G} \sigma(r, g) \qquad R(r) = \frac{1}{|R|}$$

$$= \sum_{r \in R} R(r) \cdot \sigma^*_G(r) \qquad \text{WSA-F}_1$$

- Address greedy problem of WSA-F_1

$$M_S(\ell) = \{e \mid \sigma(\ell, e) = \sigma^*_S(\ell),\ \forall e \in S\}$$

$$precision = \sum_{r \in R} \min \left(R(r), \sum_{e \in M_G(r)} G(e) \right) \cdot \sigma^*_G(r) \qquad \text{CWSA-F}_1$$

Constrained Weighted Similarity-Aligned F1

- We start with the standard F_1 measure.

$$precision = \frac{|R \cap G|}{|R|}, \; recall = \frac{|R \cap G|}{|G|}, \; F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$ F_1

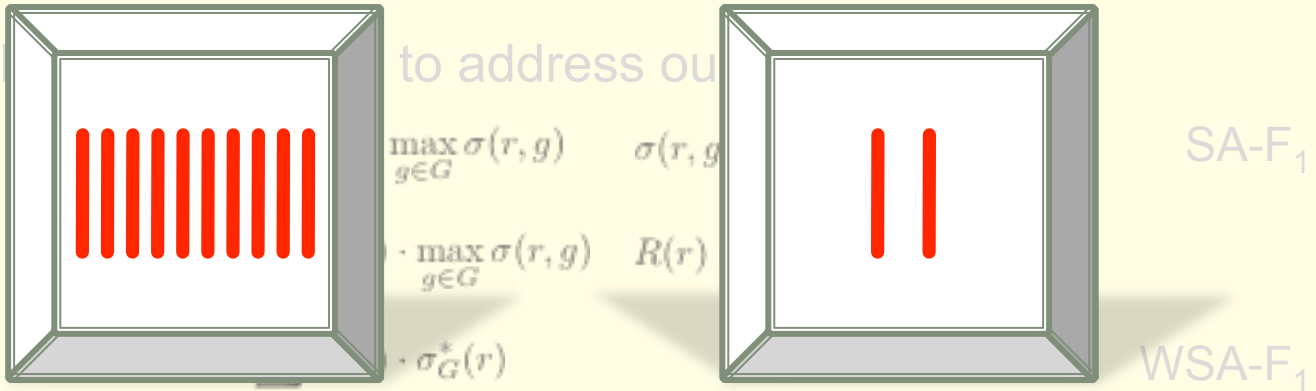
- Genera to address ou

$$\max_{g \in G} \sigma(r, g) \qquad \sigma(r, g$$

$$\cdot \max_{g \in G} \sigma(r, g) \qquad R(r)$$ SA-F_1

$$\cdot \sigma^*_G(r)$$

$$r \in R$$



WSA-F_1

- Address greedy problem of WSA-F_1

$$M_S(\ell) = \{e \mid \sigma(\ell, e) = \sigma^*_S(\ell), \; \forall e \in S\}$$

$$precision = \sum_{r \in R} \min \left(R(r), \sum_{e \in M_G(r)} G(e) \right) \cdot \sigma^*_G(r)$$ CWSA-F_1

Constrained Weighted Similarity-Aligned F1

Gold G	$(\text{angry}, 0.9), (\text{afraid}, 0.05), (\text{guilty}, 0.05)$
Response R_1	$(\text{angry}, 0.1), (\text{afraid}, 0.2), (\text{guilty}, 0.7)$
Response R_2	$(\text{irate}, 0.45), (\text{mad}, 0.45), (\text{scared}, 0.05), (\text{guilty}, 0.05)$

	F_1			SA-F_1			WSA-F_1			CWSA-F_1		
	p	r	f_1	p	r	f_1	p	r	f_1	p	r	f_1
R_1	1	1	1	1	1	1	1	1	1	0.2	0.2	0.2
R_2	0.25	$0.33\overline{3}$	0.29	1	1	1	1	1	1	1	1	1

Suppose that σ(angry, irate) = σ(angry, mad) = σ(afraid, scared) = 1,
with σ of any two identical labels being 1, and σ of all other pairs are 0.

red = non-intuitive score
green = intuitive score

Defining the Problem

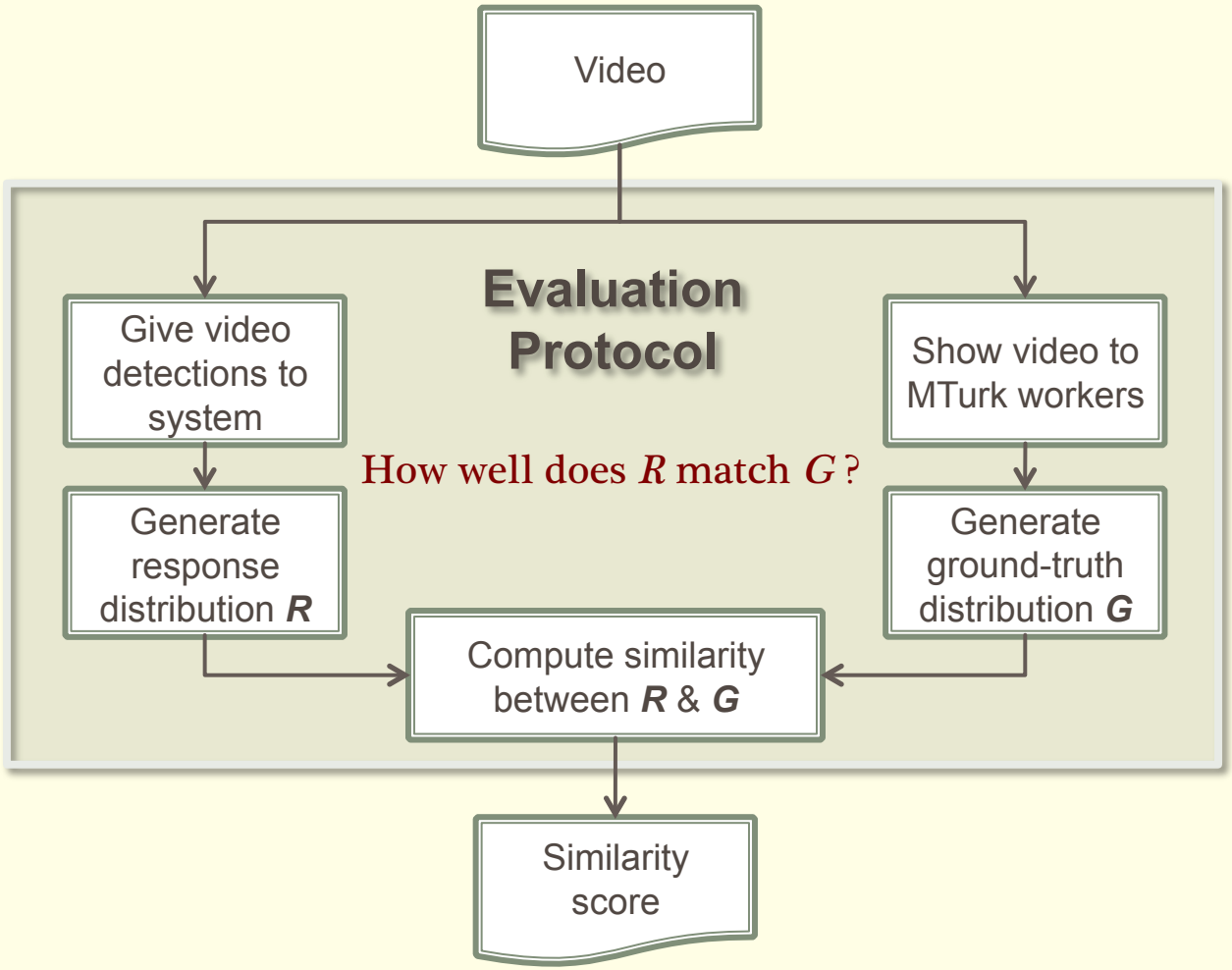
Models

Performance Measures

Results

Conclusions / Future Work

Evaluation Procedure



Mental State Identification in Chase Videos

	CWSA-F_1		
	p	r	f_1
baseline	.284	.289	.286
vector	.399	.392	.393
sentence	.366	.376	.368
coref	.382	.461	.416
event	.446	.488	.463
event+vector	**.488**	**.517**	**.500**

The average evaluation performance across 26 different chase videos are shown against the baseline scores for our neighborhood information extraction models. Bold font indicates the best score in a given column.

* All average improvements over the baseline responses are significant ($p < 0.01$). All significance tests were one-tailed and were based on nonparametric bootstrap resampling with 10,000 iterations.

Mental State Identification in Chase Videos

	CWSA-F_1		
	p	r	f_1
baseline	.284	.289	.286
vector	.399	.392	.393
sentence	.366	.376	.368
coref	.382	.461	.416
event	.446	.488	.463
event+vector	**.488**	**.517**	**.500**

- *event+vector* outperformed baseline by almost 75%.
- Ensemble outperformed individual components.
  - Operating in latent space (*vector*) and operating on text (*event*) yield complementary information.
- Incremental improvements due to NLP.
  - *sentence* < *coref* < *event*

Limitation: Biases in Underlying Data

Categories	Baseline	*event+vector*	Change
children	0.2082	0.3599	+0.1517
police	**0.3313**	**0.6006**	**+0.2693**
sports	0.2318	0.4126	+0.1808
others	0.3157	0.5457	+0.2300

The average CWSA-F_1 scores for the ensemble model *event+vector* are shown in comparison to the baseline performance, categorized by video scenarios.

children = video contains a child participant
police = video contains a policeman participant
sports = video is sports-related
other = video does not fit in the first categories (e.g., civilian adults)

Limitations: Biases in Underlying Data

Categories	Baseline	*event+vector*	Change
children	0.2082	0.3599	+0.1517
police	**0.3313**	**0.6006**	**+0.2693**
sports	0.2318	0.4126	+0.1808
others	0.3157	0.5457	+0.2300

- Baseline did worse on children and sports-related videos than police related videos.
- Baseline uses all 160 mental states with uniform probability.
- **Initial seed set is more fit to describe police chases.**
- See biggest improvement over baseline on police videos, least improvement on children videos.
- Gigaword corpus = newswire articles
- **Underlying corpus is biased towards police chases (i.e., news-worthy events).**

Actor-specific Mental State Identification

	CWSA-F_1		
	p	r	f_1
baseline	.196	.195	.195
vector	.351	.338	.342
event	.353	.340	.340
event+vector	**.395**	**.400**	**.396**

The average evaluation performance for the mental state of the **subject** across 26 different chase videos.

	CWSA-F_1		
	p	r	f_1
baseline	.191	.181	.185
vector	.358	.374	.363
event	.383	.407	.391
event+vector	**.389**	**.415**	**.399**

The average evaluation performance for the mental state of the **object** across 26 different chase videos.

Mental State Identification in Hug Videos

	CWSA-F_1		
	p	r	f_1
baseline	.226	.210	.217
vector	.347	.334	.339
sentence	.388	.378	.382
event	.406	.384	.394
event+vector	**.443**	**.437**	**.439**

The average evaluation performance across 45 different hug videos are shown against the baseline scores for our neighborhood information extraction models. Bold font indicates the best score in a given column.

Mental State Identification in Hug Videos

	CWSA-F_1		
	p	r	f_1
baseline	.226	.210	.217
vector	.347	.334	.339
sentence	.388	.378	.382
event	.406	.384	.394
event+vector	**.443**	**.437**	**.439**

- Overall lower performance
  - Data bias, only 41K documents containing hug vs. 141K documents for chase
- *event+vector* outperformed baseline by over 100%
- Consistent behavior as seen in chase videos:
  - Ensemble outperformed individual components.
  - Incremental improvement with each NLP module.

Gold G_{10}	(angry, 0.19)(scared, 0.14)(excited, 0.08)(frantic, 0.06)(fearful, 0.06), (energize, 0.03)(guilty, 0.03)(hopeless, 0.03)(panic, 0.03)(protective, 0.03) (vigilant, 0.03)(eccentric, 0.03)(frightened, 0.03)(concerned, 0.03) (determine, 0.03)(surprised, 0.03)(upset, 0.03)(crazy, 0.03) (aggressive, 0.03)(startled, 0.03)(desperate, 0.03)(focus, 0.03)
vector	(**frantic**, 0.07)(**crazy**, 0.07)(**aggressive**, 0.07)(hurry, 0.07) (**surprised**, 0.07)(bored, 0.07)(fun, 0.07)(worthless, 0.07) (**desperate**, 0.07)(furious, 0.07)(weird, 0.07)(jealous, 0.07)(nervous, 0.07) (giddy, 0.06)(**scared**, 0.06)
sentence	(**angry**, 0.11)(relax, 0.09)(calm, 0.09)(**desperate**, 0.08)(serious, 0.06) (**focus**, 0.06)(happy, 0.05)(sad, 0.04)(miserable, 0.04)(pleased, 0.04) (afraid, 0.04)(weary, 0.03)(motivated, 0.03)(energetic, 0.03)(eager, 0.02) (**concerned**, 0.02)(**determine**, 0.02)(**upset**, 0.02)(violent, 0.02) (romantic, 0.02)(**crazy**, 0.01)(**guilty**, 0.01)(reluctant, 0.01) (**aggressive**, 0.01)(cautious, 0.01)(unhappy, 0.01)(amuse, 0.01) (interested, 0.01)(worried, 0.01)(welcome, 0.01)
event+vector	(happy, 0.08)(**crazy**, 0.06)(**upset**, 0.04)(afraid, 0.04)(**determine**, 0.04) (tired, 0.04)(**frantic**, 0.04)(**aggressive**, 0.03)(hurry, 0.03) (**surprised**, 0.03)(bored, 0.03)(fun, 0.03)(worthless, 0.03) (**desperate**, 0.03)(furious, 0.03)(weird, 0.03)(jealous, 0.03)(nervous, 0.03) (giddy, 0.03)(**scared**, 0.03)(**angry**, 0.03)(**focus**, ((violent, 0.03)(**guilty**, 0.03)(disappointed, 0.03)(r (interested, 0.02)(mad, 0.02)

chase10

Defining the Problem

Models

Performance Measures

Results

Conclusions / Future Work

Conclusions

- Summary
  - **Problem**: Identifying latent attributes in videos, given some context.
  - **Data**: Videos from web, annotations via crowd sourcing
  - **Solution**: Largely unsupervised information extraction models
    - Lexical semantic in vector space (*vector*)
    - Sentence co-occurrence in text (*sentence*)
    - Event-centric in text (*event*)
  - **Evaluation**: CWSA-F_1 score to compare mental state distributions
- Findings
  - Little context needed (activity, actor-type)
  - Ensemble models perform best
  - NLP techniques help
  - Actor-specific identification is possible
  - Works for different events (e.g., *chase* and *hug*)

Future Work

- Model Improvement
  - Resolve data bias – Use ClueWeb09 corpus (contains 1 billion web pages)
  - Automatically learn new mental state labels & extraction patterns
    - Mutual bootstrapping (Riloff and Jones, 1999)
  - More contextual information? Increase size of input query tuples.
- Computer Vision Integration
  - Degrade quality of MTurk annotations to simulate noisy detectors
  - Integrate full automatic detection system
- Other Applications
  - Apply the CWSA-F1 score to other problems (e.g., image retrieval)
  - More challenging (ambiguous) verbs: *walk* and *run*
  - Medical diagnosis – identify causes relating to known context symptoms.

Acknowledgements

I would like to express my heartfelt gratitude to the following people for helping me realize my dream (in no particular order).

- My advisors.
- My committee.
- My colleagues and friends.
- **Most important of all, my family!**

THE END

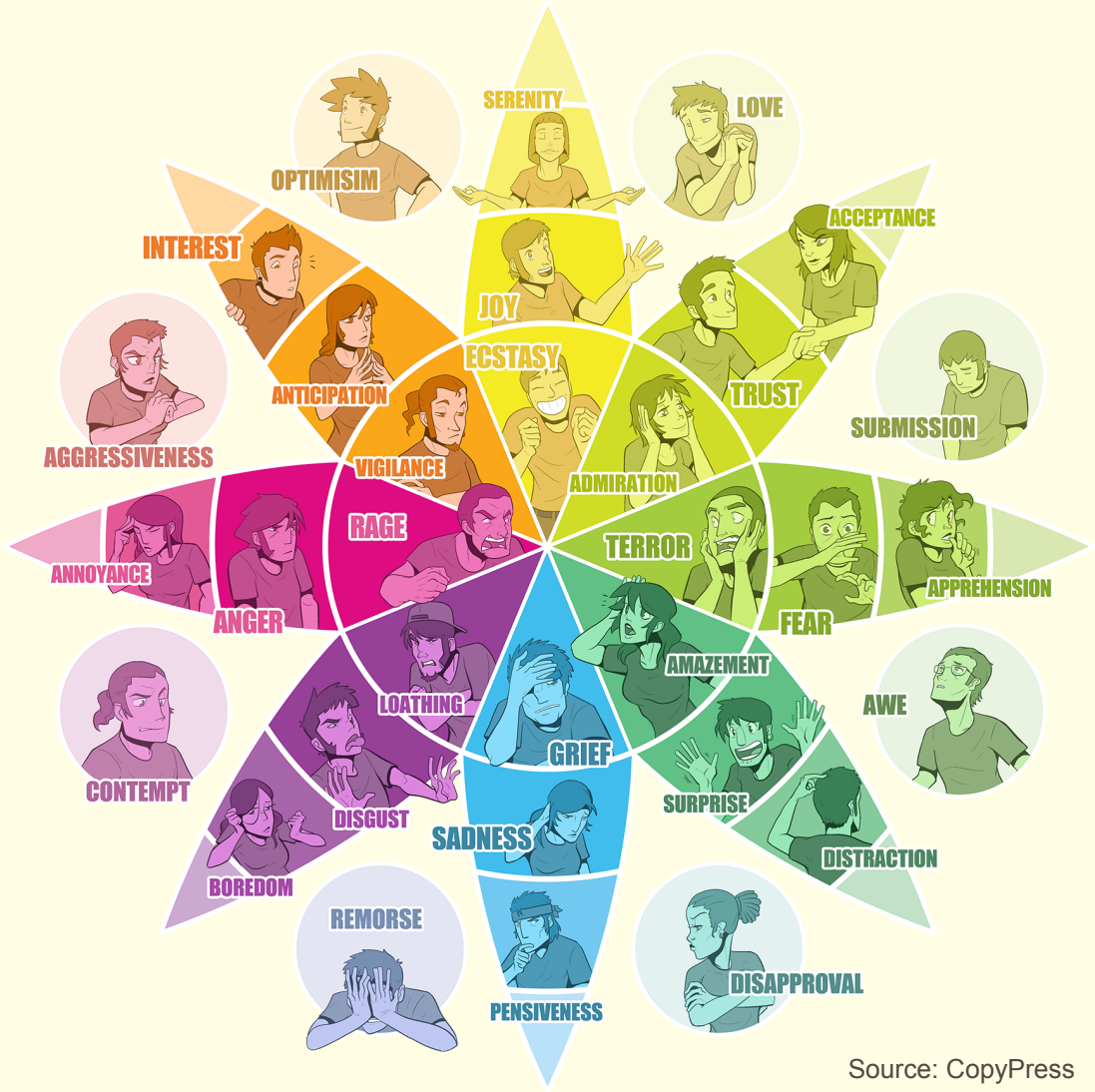
Related Work

- Mental State Inference
  - Typically formulate as a *classification* problem
    - Use low level sensory input to predict mental state labels
  - Typically modeled using HMMs and DBNs
    - Abbasi et al., 2009; Teoh and Cho, 2011; El Kaliouby and Robinson, 2004; Baltrusaitis et al., 2011
- Computer Vision
  - Wide arrays of object detectors
    - Felzenszwalb et al. 2008; Yang and Ramanan 2011, Ramanan 2012
  - Trackers 2D (Ramanan et al. 2007) and 3D (Brau et al. 2011)
  - Human action recognitions (O'Hara and Draper 2012; Sadanand and Corso 2012)
- Natural Language Processing & Information Extraction/Retrieval
  - Many text processing tools: POS tagging, syntactic dependency, named entity recognition, coreference resolution, semantic role labeling.
  - Many novel unsupervised information extraction solutions
    - Identifying meaning of "little kid" (Marneffe et al. 2010); infer yes/no answers from indirect yes/no answers (Mohtarami et al. 2011); extract narrative schemas and information templates from text (Chambers and Jurafsky, 2009, 2011)

References

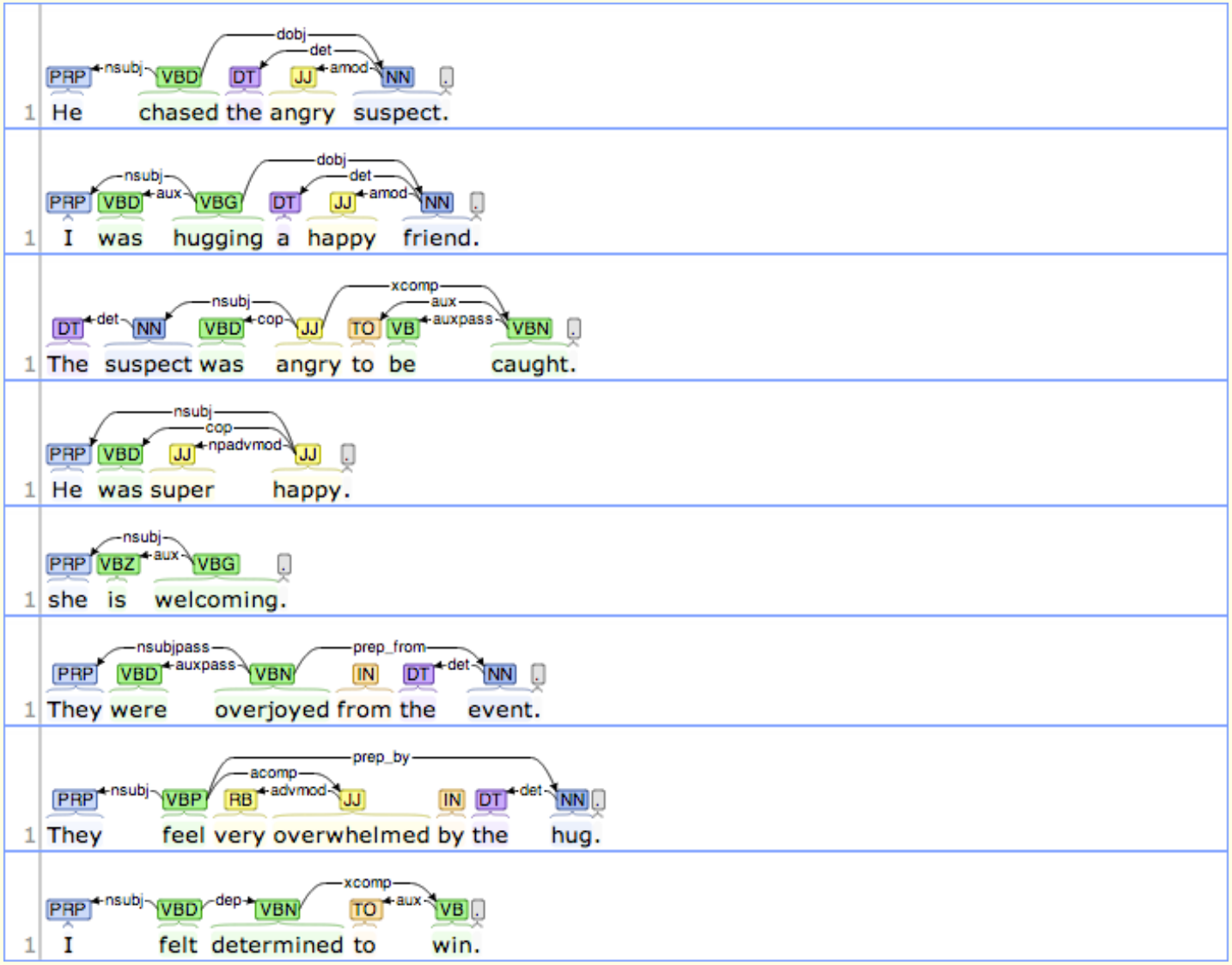
- Abbasi, Abdul Rehman et al. “Student Mental State Inference from Unintentional Body Gestures Using Dynamic Bayesian Networks.” Journal on Multimodal User Interfaces 3.1-2 (2009): 21–31. Web. 23 Jan. 2014.
- Baltrusaitis, Tadas et al. “Real-Time Inference of Mental States from Facial Expressions and Upper Body Gestures.” Face and Gesture 2011. IEEE, 2011. 909–914. Web. 23 Jan. 2014.
- Brau, Ernesto et al. “A Generative Statistical Model for Tracking Multiple Smooth Trajectories.” CVPR 2011. IEEE, 2011. 1137–1144. Web. 23 Jan. 2014.
- Chambers, Nathanael, and Dan Jurafsky. “Template-Based Information Extraction without the Templates.” Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1. N. p., 2011. 976–986. Web. 12 Apr. 2013.
- Chambers, Nathanael, and Dan Jurafsky. “Unsupervised Learning of Narrative Schemas and Their Participants.” Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP Volume 2 ACLIJCNLP 09. N. p., 2009. 602–610. Web. 11 Apr. 2013.
- El Kaliouby, R., and P. Robinson. “Real-Time Inference of Complex Mental States from Facial Expressions and Head Gestures.” 2004 Conference on Computer Vision and Pattern Recognition Workshop. IEEE, 2004. 154–154. Web. 23 Jan. 2014.
- Felzenszwalb, Pedro, David McAllester, and Deva Ramanan. “A Discriminatively Trained, Multiscale, Deformable Part Model.” 2008 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2008. 1–8. Web. 23 Jan. 2014.
- Marneffe, MC De, CD Manning, and Christopher Potts. “‘Was It Good? It Was Provocative.’ Learning the Meaning of Scalar Adjectives.” Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. N. p., 2010. 167–176. Web. 15 Apr. 2013.
- Mikolov, Tomas et al. “Efficient Estimation of Word Representations in Vector Space.” arXiv preprint arXiv:1301.3781 (2013): 1–12. Web. 26 Aug. 2013.
- Mohtarami, Mitra et al. “Predicting the Uncertainty of Sentiment Adjectives in Indirect Answers.” Proceedings of the 20th ACM International Conference on Information and Knowledge Management - CIKM ’11. New York, New York, USA: ACM Press, 2011. 2485. Web. 23 Jan. 2014.
- O’Hara, S, and B. A. Draper. “Scalable Action Recognition with a Subspace Forest.” 2012 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2012. 1210–1217. Web. 23 Jan. 2014.
- Ramanan, D. “Face Detection, Pose Estimation, and Landmark Localization in the Wild.” 2012 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2012. 2879–2886. Web. 23 Jan. 2014.
- Ramanan, Deva, David a Forsyth, and Andrew Zisserman. “Tracking People by Learning Their Appearance.” IEEE transactions on pattern analysis and machine intelligence 29.1 (2007): 65–81. Web. 23 Jan. 2014.
- Riloff, E, and R Jones. “Learning Dictionaries for Information Extraction by Multi-Level Bootstrapping.” Proceedings of the Sixteenth National Conference on Artificial Intelligence (AAAI-1999). N. p., 1999. 474–479. Web. 15 July 2013.
- Rubner, Yossi, Carlo Tomasi, and LJ Guibas. “The Earth Mover’s Distance as a Metric for Image Retrieval.” International Journal of Computer Vision 40.2 (2000): 99–121. Web. 26 Apr. 2014.
- Sadanand, S., and J. J. Corso. “Action Bank: A High-Level Representation of Activity in Video.” 2012 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2012. 1234–1241. Web. 23 Jan. 2014.
- Teoh, Teik-Toe, and Siu-Yeung Cho. “Human Emotional States Modeling by Hidden Markov Model.” 2011 Seventh International Conference on Natural Computation. IEEE, 2011. 908–912. Web. 23 Jan. 2014.
- Yang, Yi, and Deva Ramanan. “Articulated Pose Estimation with Flexible Mixtures-of-Parts.” CVPR 2011. IEEE, 2011. 1385–1392. Web. 23 Jan. 2014.

BACKUP SLIDES



Source: CopyPress

admiring
afraid
aggressive
agitated
alarmed
alert
ambitious
amazed
amused
angry
annoyed
anxious
apprehensive
ashamed
assertive
bitter
bored
calm
carefree
cautious
cheerful
competitive
complacent
concerned
confused
considerate
content
cranky
crazy
curious
cynical
demented
depressed
desperate
determined
devious
disappointed
discontent
discouraged
disgusted
distracted
drunken
eager
ecstatic
encouraged
energetic
energized
enraged
enthusiastic
envious
excited
exhausted
exhilarating
fatigued
fearful
focused
forgetful
frantic
friendly
frightened
frustrated
fun
furious
giddy
glamorous
gleeful
grateful
grumpy
guilty
happy
helpful
helpless
homicidal
hopeful
hopeless
hostile
hurried
impressed
indifferent
inebriated
infuriated
instinctive
interested
irate
irritated
jealous
joyful
joyous
lively
loathsome
lonely
loved
mad
mellow
merciless
mischievous
miserable
motivated
naughty
nervous
numb
optimistic
panicked
panicky
peaceful
peeved
pessimistic
playful
pleased
protective
raging
rebellious
refreshed
relaxed
relieved
reluctant
remorseful
resentful
restless
revengeful
romantic
sad
satisfied
scared
selfish
selfless
serious
shaky
shocked
sickened
spiteful
stressed
submissive
surprised
suspenseful
sympathetic
tense
terrified
terrifying
thankful
threatening
tired
trustful
trusting
uncomfortab
uneasy
unhappy
unworthy
upset
urgent
vengeful
vigilant
vigorous
violent
wary
weary
weird
welcoming
worried
worthless


PRP nsubj VBD dobj DT det JJ amod NN .
1 He chased the angry suspect.
PRP nsubj VBD aux VBG dobj DT det JJ amod NN .
1 I was hugging a happy friend.
DT det NN nsubj VBD cop JJ xcomp TO aux VB auxpass VBN .
1 The suspect was angry to be caught.
PRP nsubj VBD cop JJ npadvmod JJ .
1 He was super happy.
PRP nsubj VBZ aux VBG .
1 she is welcoming.
PRP nsubjpass VBD auxpass VBN prep_from IN DT det NN .
1 They were overjoyed from the event.
PRP nsubj VBP acomp RB advmod JJ prep_by IN DT det NN .
1 They feel very overwhelmed by the hug.
PRP nsubj VBD dep VBN xcomp TO aux VB .
1 I felt determined to win.


1 A man chases a dog.
1 A man hugged his wife.
1 The owner was hugging his wife.
1 A man frightened the dog to chase it .
1 A man was chased by the police.
1 She was hugged by her mom.
1 The police chased after the suspect.
1 The man was chasing after the thief.

participants-nlp-3

```
1  ==== DOCUMENT ====
2  The police chased John around the corner . John was very afraid .
3
4
5  ==> Step 1: PARTICIPANTS IDENTIFICATION
6
7  Target Sentence(s):
8  "The police chased John around the corner ."
9
10 Identified subject (CoreNLP):   "police"
11 Identified object (CoreNLP):    "John"
12
13
14 ==> Step 2: COREFERENCE RESOLUTION
15
16 Coreference chain(s) for SUBJECTS:
17 One chain found containing the following mentions:
18     sentence (0):  [The police] chased John around the corner .
19
20 Coreference chain(s) for OBJECTS:
21 One chain found containing the following mentions:
22     sentence (1):  [John] was very afraid .
23     sentence (0): The police chased [John] around the corner .
24
25
26 ==> Step 3: COMPLEMENTS EXTRACTION
27
28 Complement(s) for SUBJECTS:
29     sentence (0): NONE
30
31 Complement(s) for OBJECT:
32     sentence (0): NONE
33     sentence (1): afraid
```

Line: 1 Column: 1 Plain Text Tab Size: 4

Document: [A . **B** . C . **D** . E . F . G . **H** . I . J]

Models:		Neighborhoods:
sentence	A . **B** . C . **D** . E . F . G . **H** . I . J	{B} ; {D} ; {H}
win-0	A . **B** . C . **D** . E . F . G . **H** . I . J	{B, D, H}
win-1	A . **B** . C . **D** . E . F . G . **H** . I . J	{A, B, C, D, E, G, H, I}
win-2	A . **B** . C . **D** . E . F . G . **H** . I . J	{A, B, C, D, E, F, G, H, I, J}
event/coref	A . **B** . C . **D** . E . F . G . **H** . I . J	{A, B, C, D, G, H}

Letters A through J represent sentences that form a document. A bolded red letter denotes a target sentence. Arrows represent links in a coreference chain.

Mental State Identification in Chase Videos

	F_1			CWSA-F_1		
	p	r	f_1	p	r	f_1
baseline	.107	**.750**	.187	.284	.289	.286
vector	.226	.145	.175	.399	.392	.393
sentence	.194	.293	.227	.366	.376	.368
coref	**.264**	.251	.253	.382	.461	.416
event	.231	.303	.256	.446	.488	.463
event+vector	.259	.296	**.274**	**.488**	**.517**	**.500**

The average evaluation performance across 26 different chase videos are shown against the baseline scores for our neighborhood information extraction models. Bold font indicates the best score in a given column.

* All average improvements over the baseline responses are significant ($p < 0.01$). All significance tests were one-tailed and were based on nonparametric bootstrap resampling with 10,000 iterations.

Effectiveness of Coreference Resolution

Models	CWSA-F1	Versus *coref*	*p*-value
win-0	0.388682	−0.027512	0.0067
win-1	0.415328	−0.000866	0.4629
win-2	0.399777	−0.016417	0.0311
win-3	0.392832	−0.023362	0.0029

Comparing the average CWSA-F1 scores of a naïve windowing model, under different window sizes, to the performance of the *coref* model. The *p*-values, based on the average differences, were obtained using one-tailed nonparametric bootstrap resampling with 10,000 iterations.

win-n extends the single sentence boundary of *sentence* to also include the *n* preceeding and *n* following sentences, while also piecing all relevant sentences of a document together to generate 1 neighborhood per document.

Ensemble Models

	F_1			CWSA-F_1		
	p	r	f_1	p	r	f_1
vector	.226	.145	.175	.399	.392	.393
sentence	.194	.293	.227	.366	.376	.368
sentence+vector	.192	.377	.250	.434	.444	.438
coref	.264	.251	.253	.382	.461	.416
coref+vector	.231	.337	.271	.448	.481	.462
event	.231	.303	.256	.446	.488	.463
event+vector	.259	.296	.274	.488	.517	.500

- Combine a deleted interpolation model (text space) with vector model (latent space) creates an ensemble model.
- Every ensemble model outperformed its respective individual components.
- **Information gained from operating on text and operating in the latent vector space are highly complementary.**
  - Improvement to each will improve the resulting ensemble model.

Effectiveness of Coreference Resolution

Models	CWSA-F1	Versus *coref*	*p*-value
win-0	0.388682	−0.027512	0.0067
win-1	0.415328	−0.000866	0.4629
win-2	0.399777	−0.016417	0.0311
win-3	0.392832	−0.023362	0.0029

- *coref* outperformed all tested windowing configurations.
- Improvement over *win-1* is not significant.
  - *coref* and *win-1* generate very similar neighborhoods (extracted roughly the same number of sentences relevant to *chase*).
- ***coref* does not do worse + provides references to participants for downstream process.**

Models	Total Sentences
win-0	90,399
win-1	260,423
coref	281,666
win-2	418,827
win-3	567,706

Semantic role labeling

	F_1			CWSA-F_1		
	p	r	f_1	p	r	f_1
event	.231	.303	.256	.446	.488	.463
event-srl	.232	.306	.259	.437	.487	.458

Description	Count
Number of documents parsed with CoreNLP	82
Number of documents parsed with SwiRL	82
Number of relevant *chase* instances	59
Number of times CoreNLP found both participants	29
Number of times SwiRL found both participants	48
Number of times CoreNLP failed to find at least one participant and SwiRL found both	20

Description	Count
Number of documents parsed with CoreNLP	141,875
Number of documents parsed with SwiRL	41,725
Number of relevant *chase* instances	91,167
Number of times CoreNLP found both participants	40,145
Number of times SwiRL found both participants	22,001
Number of times CoreNLP failed to find at least one participant and SwiRL found both	10,056

Vectors Combination Operators

	F_1			CWSA-F_1		
	p	r	f_1	p	r	f_1
vector	.226	.145	.175	.399	.392	.393
vector-mult	.162	.103	.125	.333	.343	.335

Gold G_{23}	(happy, 0.31)(playful, 0.27)(excited, 0.06)(joyful, 0.06)(adrenalin, 0.04) (fun, 0.04)(laughable, 0.04)(entice, 0.02)(wary, 0.02)(enthusiastic, 0.02) (lead, 0.02)(ecstatic, 0.02)(enjoyable, 0.02)(intrigued, 0.02) (delighted, 0.02)(aggressive, 0.02)
vector	(frantic, 0.07)(crazy, 0.07)(**aggressive**, 0.07)(surprised, 0.07)(hurry, 0.07) (bored, 0.07)(**fun**, 0.07)(desperate, 0.07)(furious, 0.07)(giddy, 0.07) (worthless, 0.07)(nervous, 0.07)(weird, 0.07)(scared, 0.06)(jealous, 0.06)
sentence	(determine, 0.07)(angry, 0.06)(focus, 0.05)(**aggressive**, 0.04) (serious, 0.04)(worried, 0.04)(**happy**, 0.03)(reluctant, 0.03) (desperate, 0.03)(violent, 0.03)(interested, 0.03)(distract, 0.03) (nervous, 0.03)(alert, 0.02)(eager, 0.02)(upset, 0.02)(welcome, 0.02) (friendly, 0.02)(romantic, 0.02)(frustrated, 0.02)(exhaust, 0.02) (frightened, 0.02)(crazy, 0.02)(terrify, 0.02)(afraid, 0.01) (competitive, 0.01)(curious, 0.01)(cautious, 0.01)(tired, 0.01)(frantic, 0.01) (bitter, 0.01)(furious, 0.01)(content, 0.01)(tense, 0.01)(irate, 0.01) (peaceful, 0.01)(weary, 0.01)(threatening, 0.01)(disappointed, 0.01) (guilty, 0.01)(relax, 0.01)(calm, 0.01)(**wary**, 0.01)(drunken, 0.01) (enraged, 0.01)(optimistic, 0.01)(hopeful, 0.01)(mad, 0.01)(hostile, 0.01) (rage, 0.01)(surprised, 0.01)(anxious, 0.01)(ambitious, 0.01)(lively, 0.01) (concerned, 0.01)
coref	(optimistic, 0.15)(protective, 0.13)(**happy**, 0.10)(guilty, 0.08) (determine, 0.07)(mad, 0.07)(focus, 0.05)(serious, 0.05)(curious, 0.04) (angry, 0.03)(scared, 0.03)(rage, 0.03)(afraid, 0.03)(welcome, 0.03) (crazy, 0.03)(nervous, 0.03)(discourage, 0.02)(violent, 0.02)
event+vector	(optimistic, 0.10)(scared, 0.09)(**happy**, 0.07)(nervous, 0.07) (guilty, 0.06)(determine, 0.05)(afraid, 0.04)(frantic, 0.04)(crazy, 0 (**aggressive**, 0.03)(surprised, 0.03)(hurry, 0.03)(bored, 0.03)(**fun** (desperate, 0.03)(furious, 0.03)(focus, 0.03)(giddy, 0.03)(worthles (weird, 0.03)(jealous, 0.03)(rage, 0.03)(mad, 0.02)

chase23