# Using Perception to Supervise Language Learning and Language to Supervise Perception

Ray Mooney

Department of Computer Sciences

University of Texas at Austin

Joint work with

David Chen, Sonal Gupta,

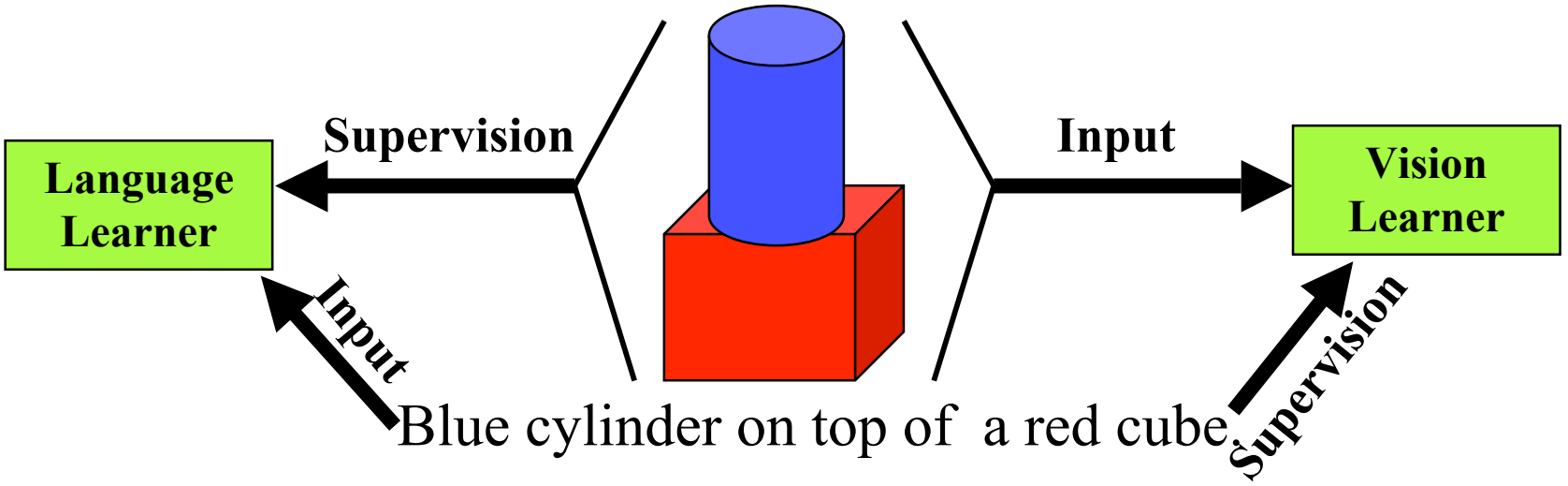
Joohyun Kim, Rohit Kate, Kristen Grauman

# Learning for Language and Vision

- Natural Language Processing (NLP) and Computer Vision (CV) are both very challenging problems.
- Machine Learning (ML) is now extensively used to automate the construction of both effective NLP and CV systems.
- Generally uses supervised ML and requires difficult and expensive human annotation of large text or image/video corpora for training.

# Cross-Supervision of Language and Vision

- Use naturally co-occurring perceptual input to supervise language learning.
- Use naturally co-occurring linguistic input to supervise visual learning.



# Using Perception to Supervise Language: Learning to Sportscast

(Chen & Mooney, ICML-08)

# Semantic Parsing

- A ***semantic parser*** maps a natural-language sentence to a complete, detailed semantic representation: ***logical form*** or ***meaning representation*** (***MR***).

- For many applications, the desired output is immediately executable by another program.

- Sample test application:
  - CLang: RoboCup Coach Language

# **CLang**: RoboCup **Coach** **Lang**uage

- In RoboCup Coach competition teams compete to coach simulated soccer players
- The coaching instructions are given in a formal language called CLang

Coach

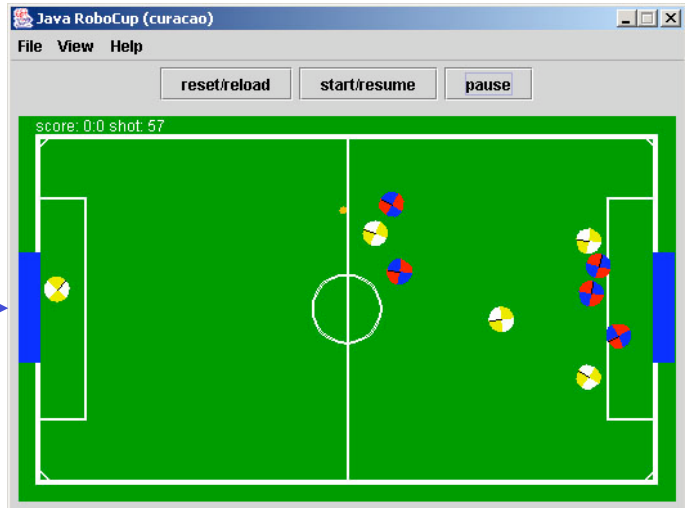
**If the ball is in our penalty area, then all our players except player 4 should stay in our half.**

Semantic Parsing

CLang

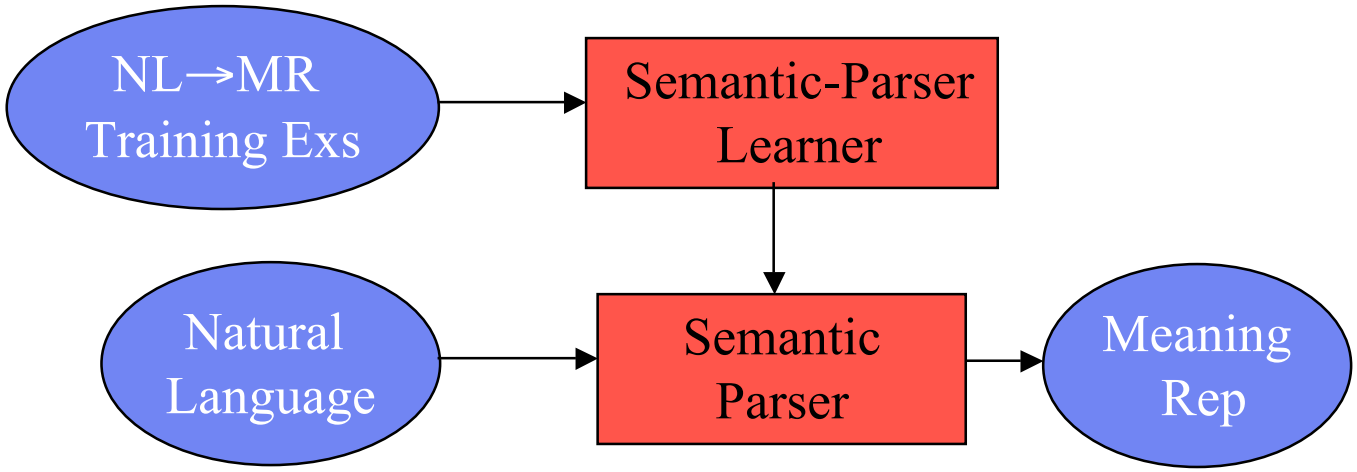
```
((bpos (penalty-area our))
(do (player-except our{4}) (pos (half our)))
```

Simulated soccer field



# Learning Semantic Parsers

- Manually programming robust semantic parsers is difficult due to the complexity of the task.
- Semantic parsers can be learned automatically from sentences paired with their logical form.



# Our Semantic-Parser Learners

- CHILL+WOLFIE (Zelle & Mooney, 1996; Thompson & Mooney, 1999, 2003)
  - Separates parser-learning and semantic-lexicon learning.
  - Learns a deterministic parser using ILP techniques.
- COCKTAIL (Tang & Mooney, 2001)
  - Improved ILP algorithm for CHILL.
- SILT (Kate, Wong & Mooney, 2005)
  - Learns symbolic transformation rules for mapping directly from NL to MR.
- SCISSOR (Ge & Mooney, 2005)
  - Integrates semantic interpretation into Collins' statistical syntactic parser.
- ⇒ WASP (Wong & Mooney, 2006; 2007)
  - Uses syntax-based statistical machine translation methods.
- ⇒ KRISP (Kate & Mooney, 2006)
  - Uses a series of SVM classifiers employing a string-kernel to iteratively build semantic representations.



# WASP

## A Machine Translation Approach to Semantic Parsing

- Uses latest statistical machine translation techniques:
  - Synchronous context-free grammars (SCFG) (Wu, 1997; Melamed, 2004; Chiang, 2005)
  - Statistical word alignment (Brown et al., 1993; Och & Ney, 2003)
- SCFG supports both:
  - Semantic Parsing: NL → MR
  - Tactical Generation: MR → NL



# KRISP

## A String Kernel/SVM Approach to Semantic Parsing

- Productions in the formal grammar defining the MR are treated like semantic concepts.
- An SVM classifier is trained for each production using a ***string subsequence kernel*** (Lodhi et al.,2002) to recognize phrases that refer to this concept.
- Resulting set of string classifiers is used with a version of Early's CFG parser to compositionally build the most probable MR for a sentence.

# Learning Language from Perceptual Context

- Children do not learn language from annotated corpora.
- Neither do they learn language from just reading the newspaper, surfing the web, or listening to the radio.
  - Unsupervised language learning
  - DARPA Learning by Reading Program
- The natural way to learn language is to perceive language in the context of its use in the physical and social world.
- This requires inferring the meaning of utterances from their perceptual context.

# Ambiguous Supervision for Learning Semantic Parsers

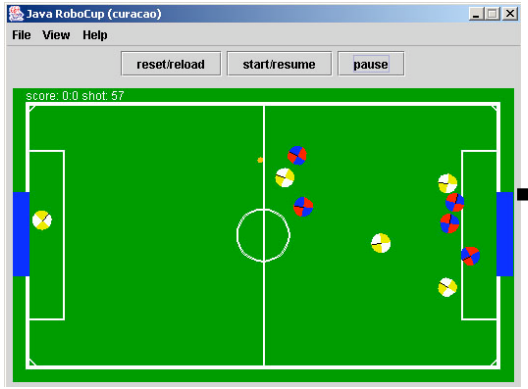
- A computer system simultaneously exposed to perceptual contexts and natural language utterances should be able to learn the underlying language semantics.
- We consider ***ambiguous*** training data of sentences associated with multiple potential MRs.
  - Siskind (1996) uses this type "referentially uncertain" training data to learn meanings of words.
- Extracting meaning representations from perceptual data is a difficult unsolved problem.
  - Our system directly works with symbolic MRs.

# Tractable Challenge Problem: Learning to Be a Sportscaster

- **Goal**: Learn from realistic data of natural language used in a representative context while avoiding difficult issues in computer perception (i.e. speech and vision).
- **Solution**: Learn from textually annotated traces of activity in a simulated environment.
- **Example**: Traces of games in the Robocup simulator paired with textual sportscaster commentary.

# Grounded Language Learning in Robocup

## Robocup Simulator




Simulated Perception

Perceived Facts

## Sportscaster



Score!!!!

Grounded Language Learner

Language Generator

SCFG

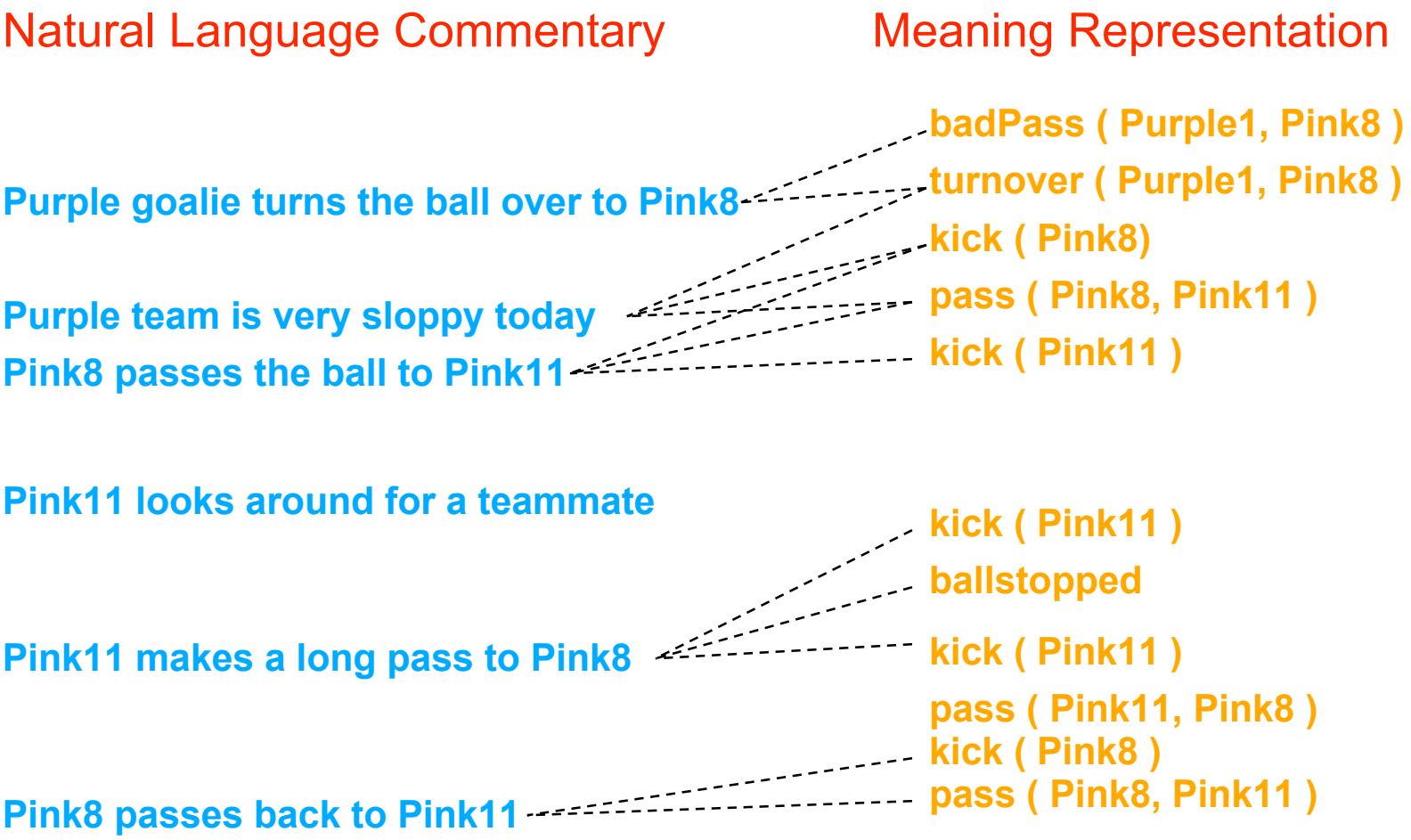
Score!!!!

Semantic Parser

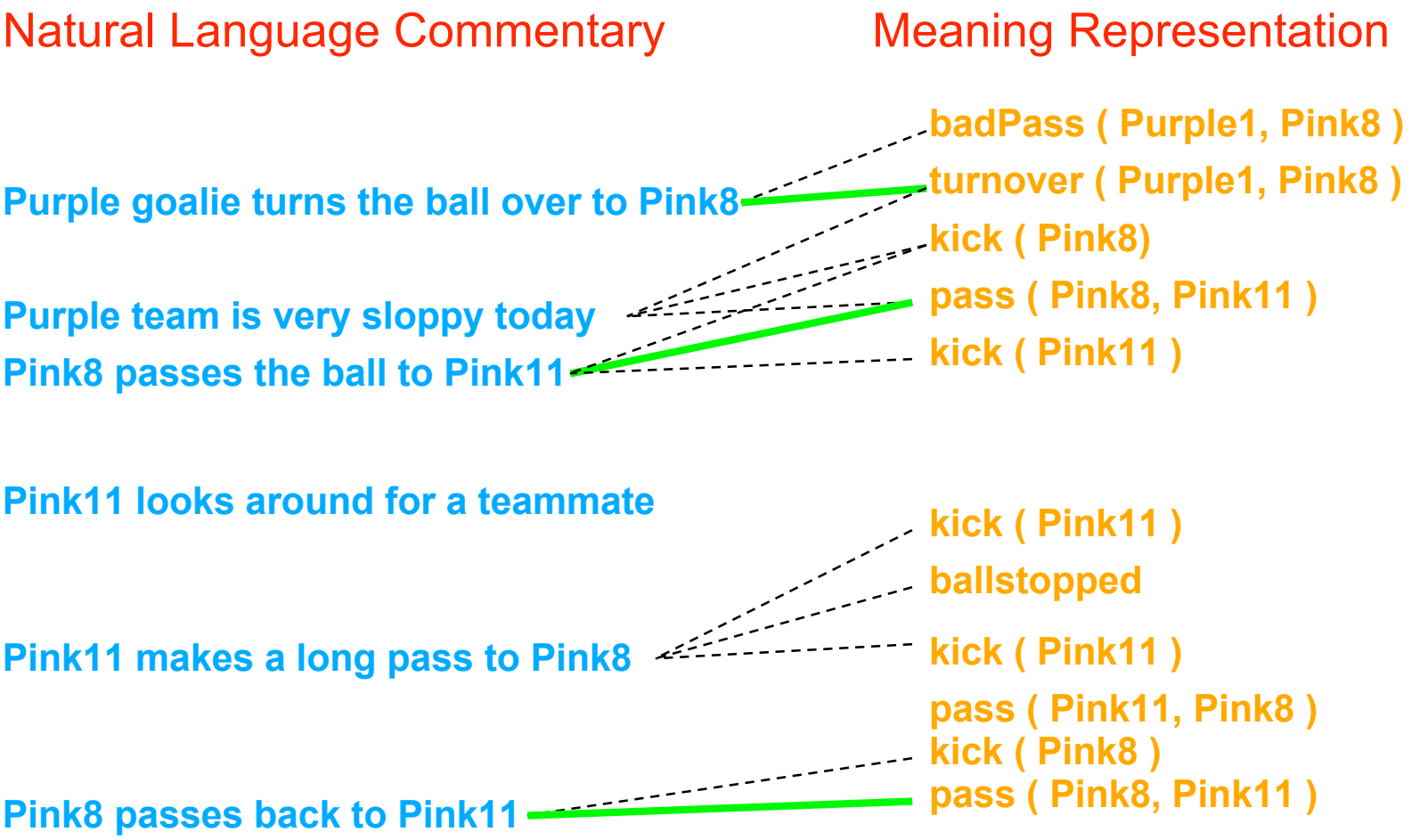
# Robocup Sportscaster Trace

| Natural Language Commentary | Meaning Representation |
| --- | --- |
| Purple goalie turns the ball over to Pink8 | badPass ( Purple1, Pink8 )<br>turnover ( Purple1, Pink8 ) |
| Purple team is very sloppy today | kick ( Pink8) |
| Pink8 passes the ball to Pink11 | pass ( Pink8, Pink11 )<br>kick ( Pink11 ) |
| Pink11 looks around for a teammate | kick ( Pink11 )<br>ballstopped |
| Pink11 makes a long pass to Pink8 | kick ( Pink11 )<br>pass ( Pink11, Pink8 ) |
| Pink8 passes back to Pink11 | kick ( Pink8 )<br>pass ( Pink8, Pink11 ) |

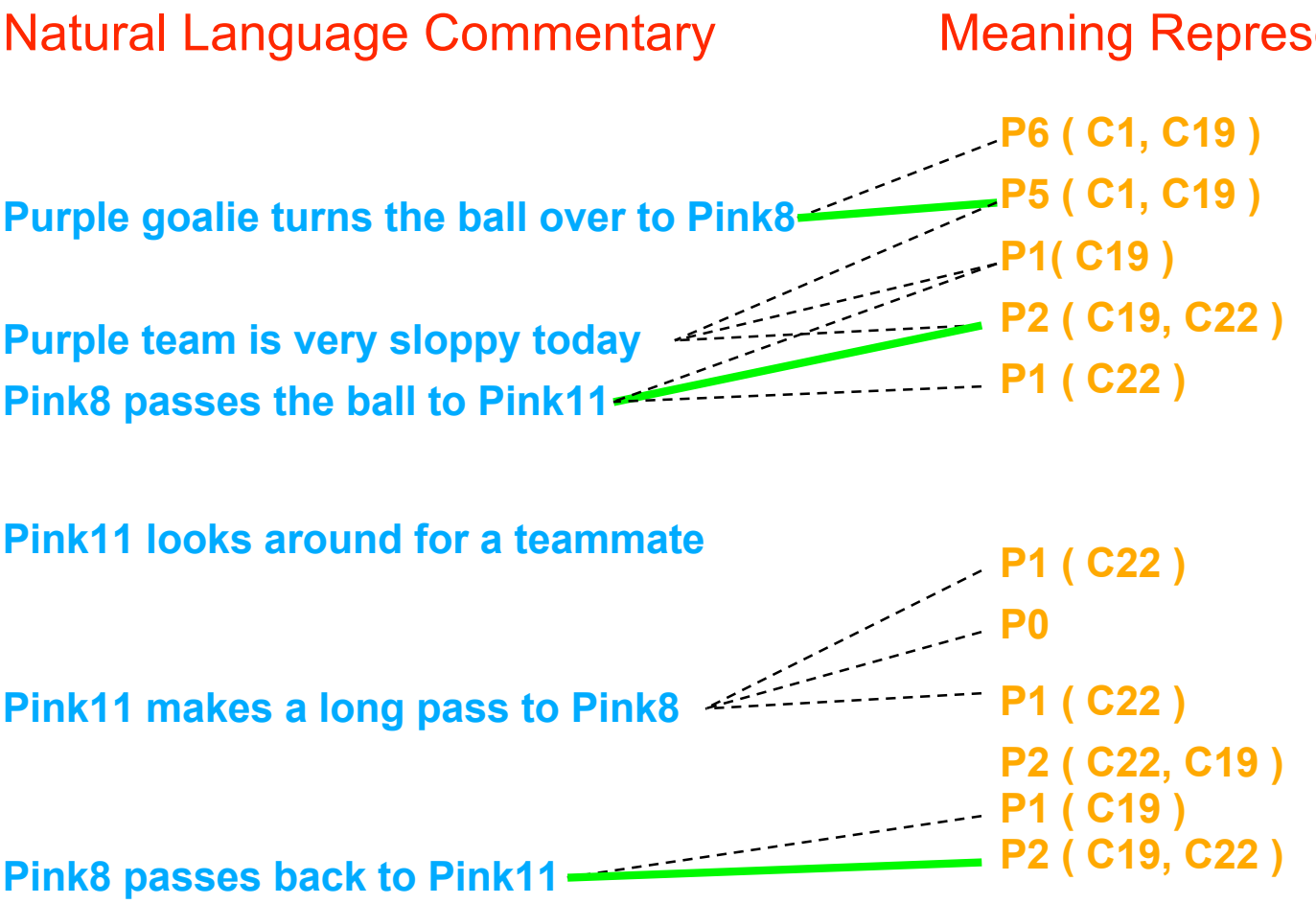
# Robocup Sportscaster Trace



# Robocup Sportscaster Trace



# Robocup Sportscaster Trace



# Sportscasting Data

- Collected human textual commentary for the 4 Robocup championship games from 2001-2004.
  - Avg # events/game = 2,613
  - Avg # sentences/game = 509
- Each sentence matched to all events within previous 5 seconds.
  - Avg # MRs/sentence = 2.5 (min 1, max 12)
- Manually annotated with correct matchings of sentences to MRs (for evaluation purposes only).

# KRISPER:
# KRISP with **E**M-like **R**etraining

- Extension of KRISP that learns from ***ambiguous*** supervision (Kate & Mooney, AAAI-07).

- Uses an iterative EM-like self-training method to gradually converge on a correct meaning for each sentence.

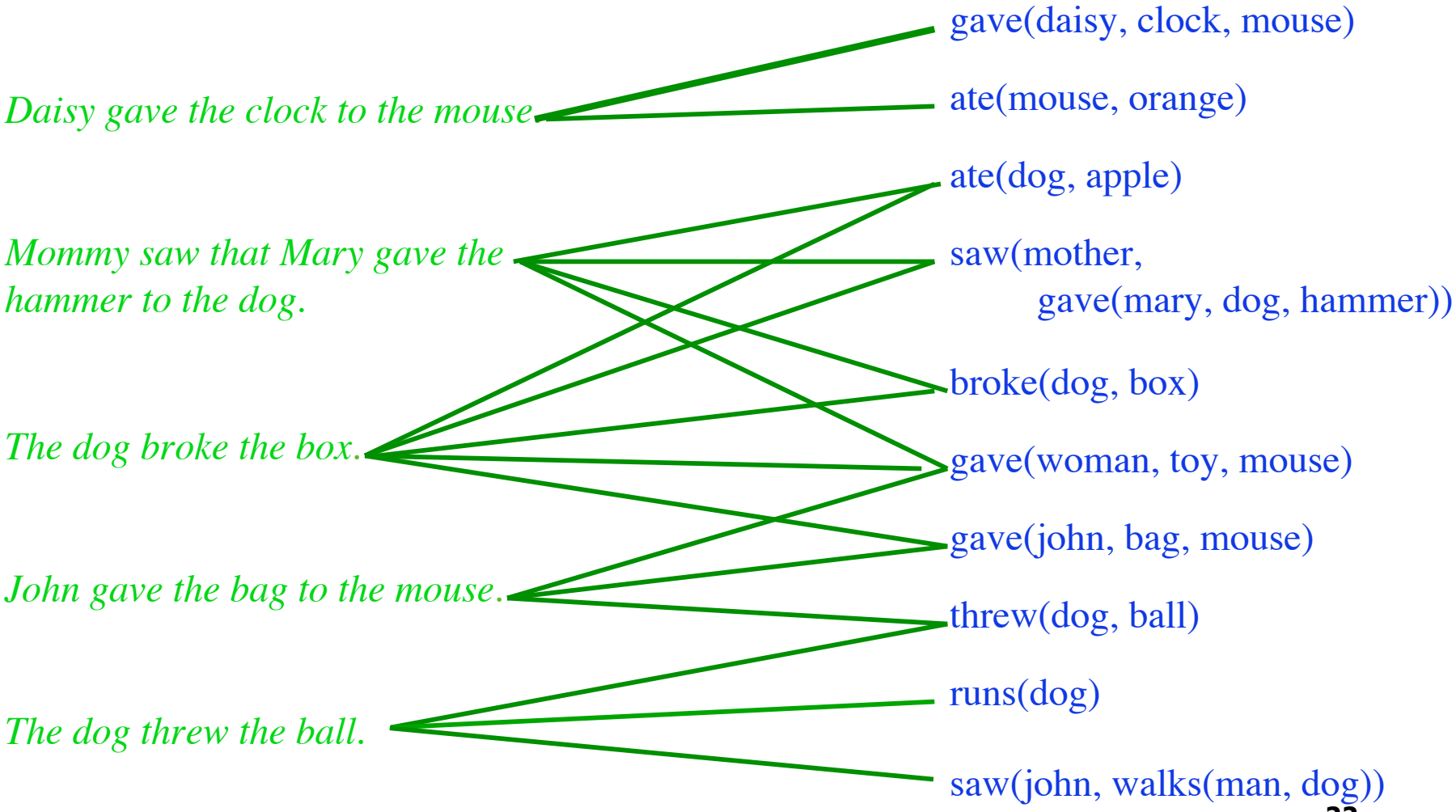
# KRISPER's Training Algorithm

1. Assume every possible meaning for a sentence is correct



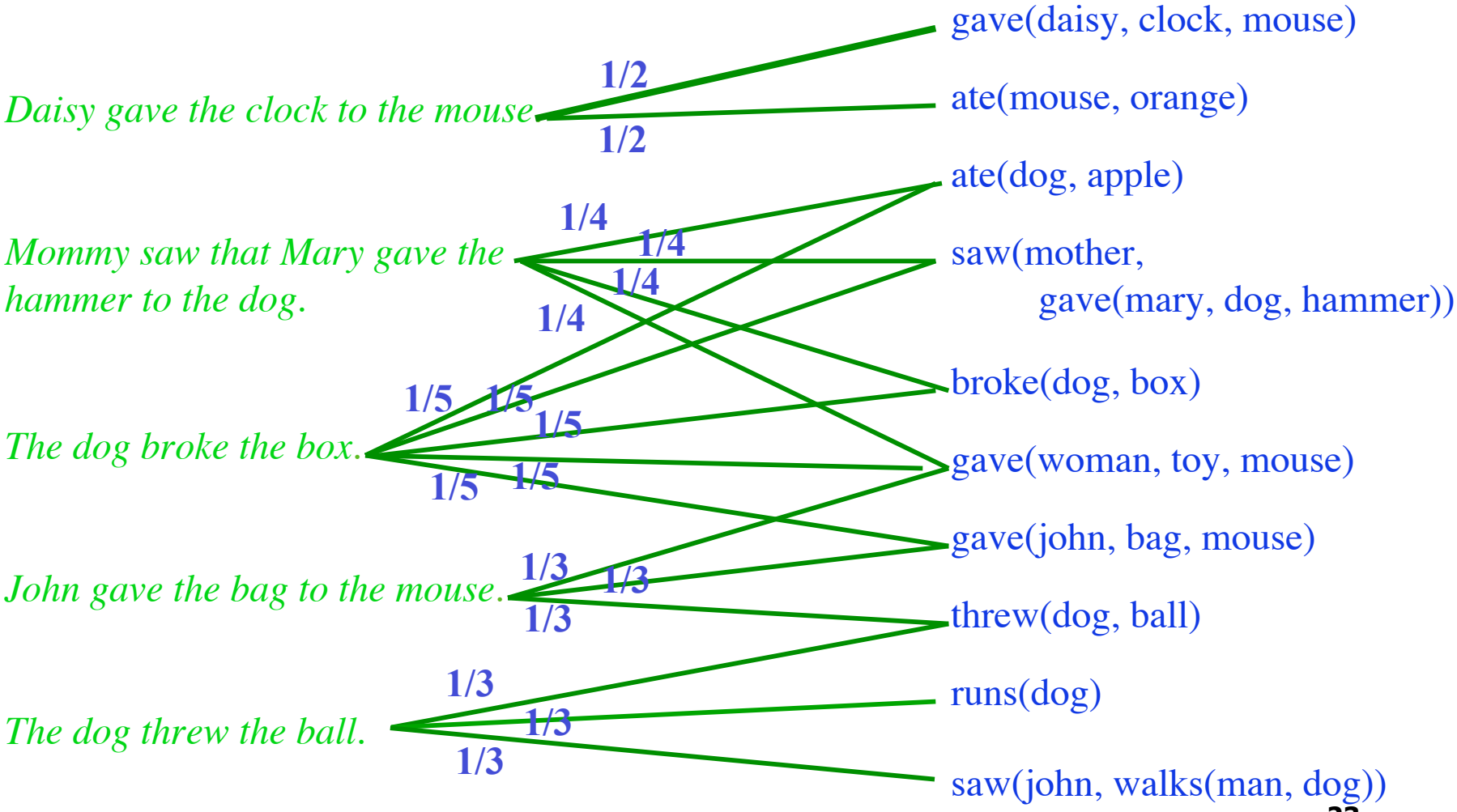
# KRISPER's Training Algorithm

1. Assume every possible meaning for a sentence is correct



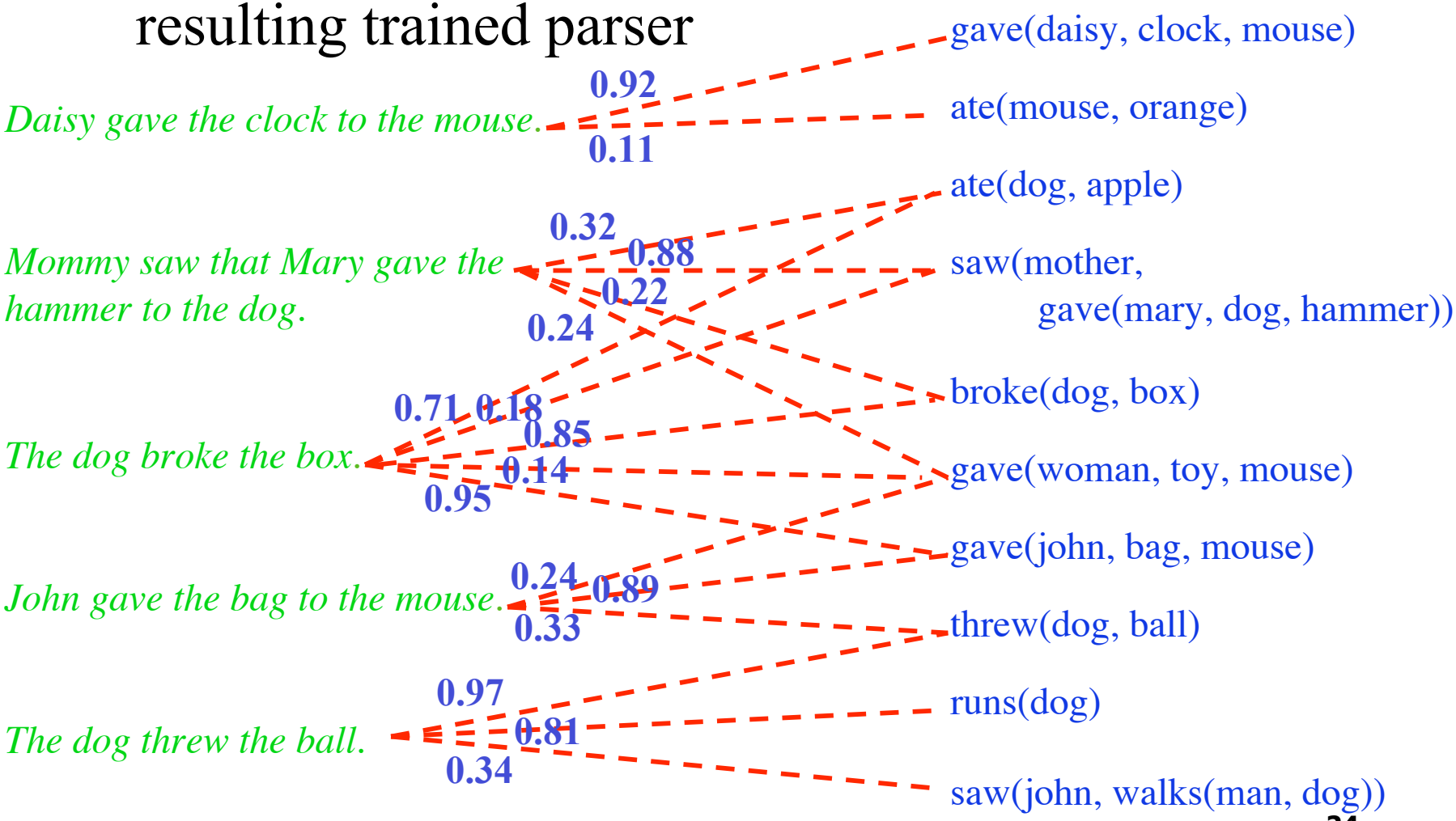
# KRISPER's Training Algorithm

## 2. Resulting NL-MR pairs are weighted and given to KRISP



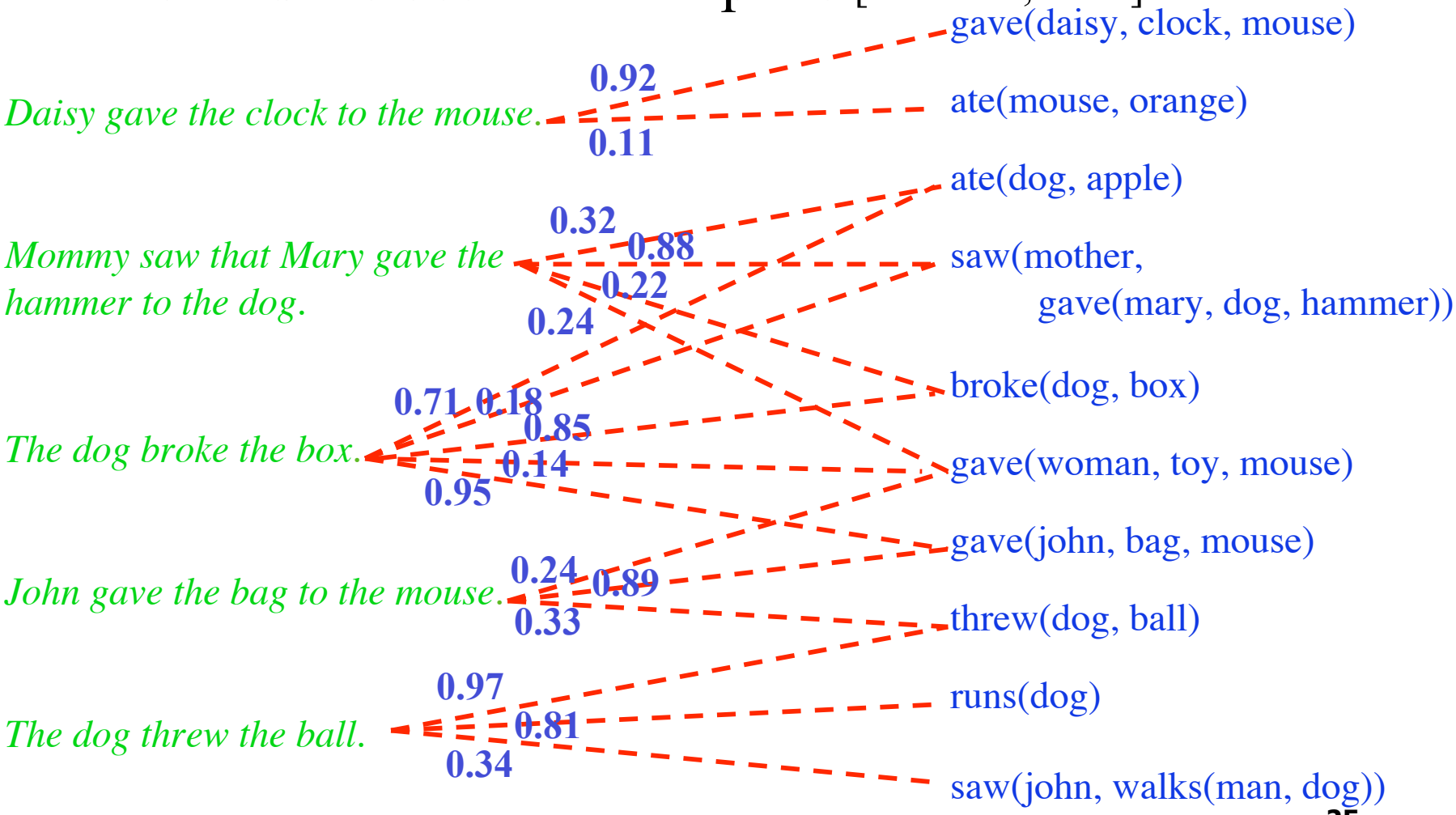
# KRISPER's Training Algorithm

3. Estimate the confidence of each NL-MR pair using the resulting trained parser



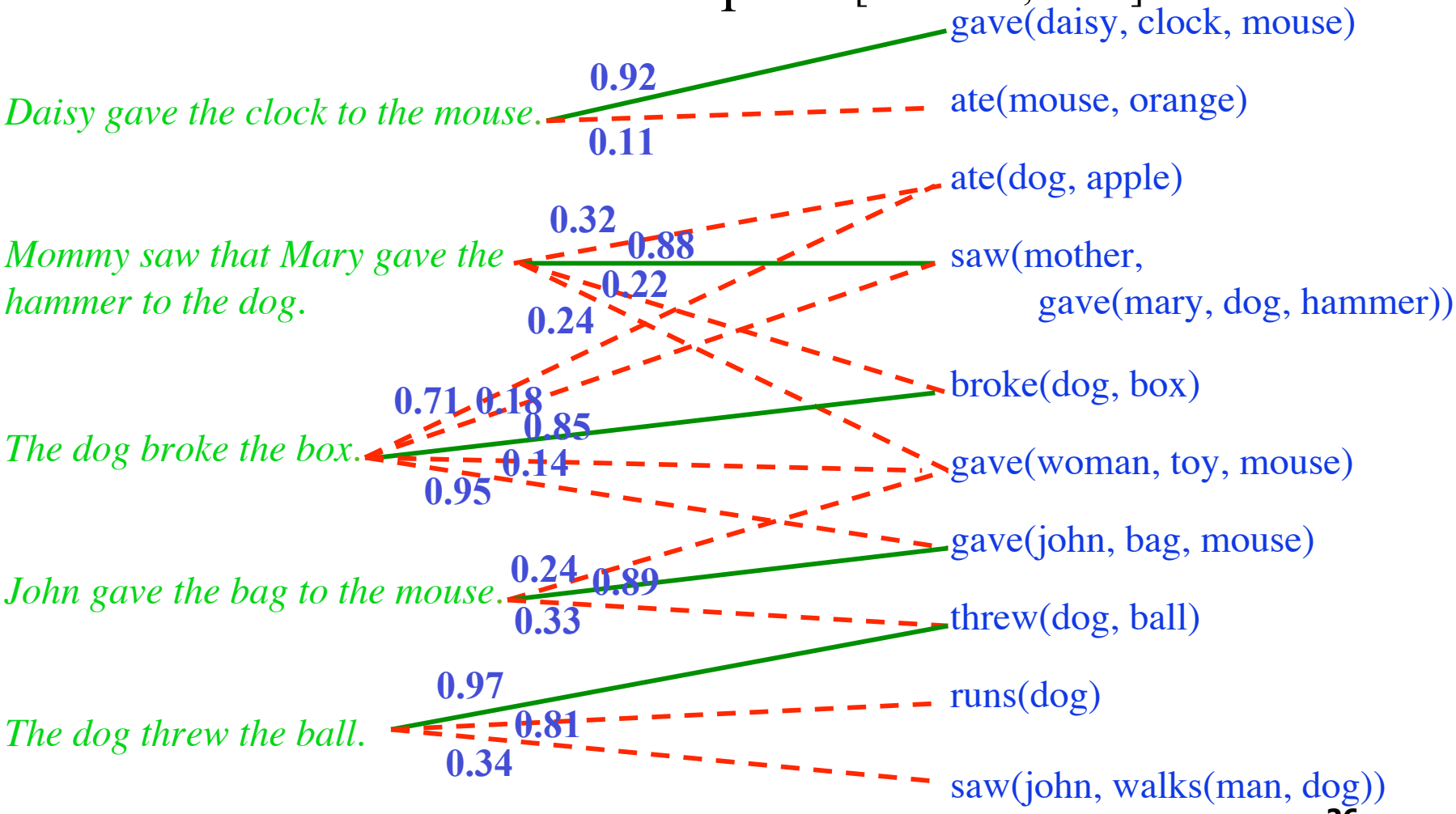
# KRISPER's Training Algorithm

4. Use *maximum weighted matching* on a bipartite graph to find the best NL-MR pairs [Munkres, 1957]



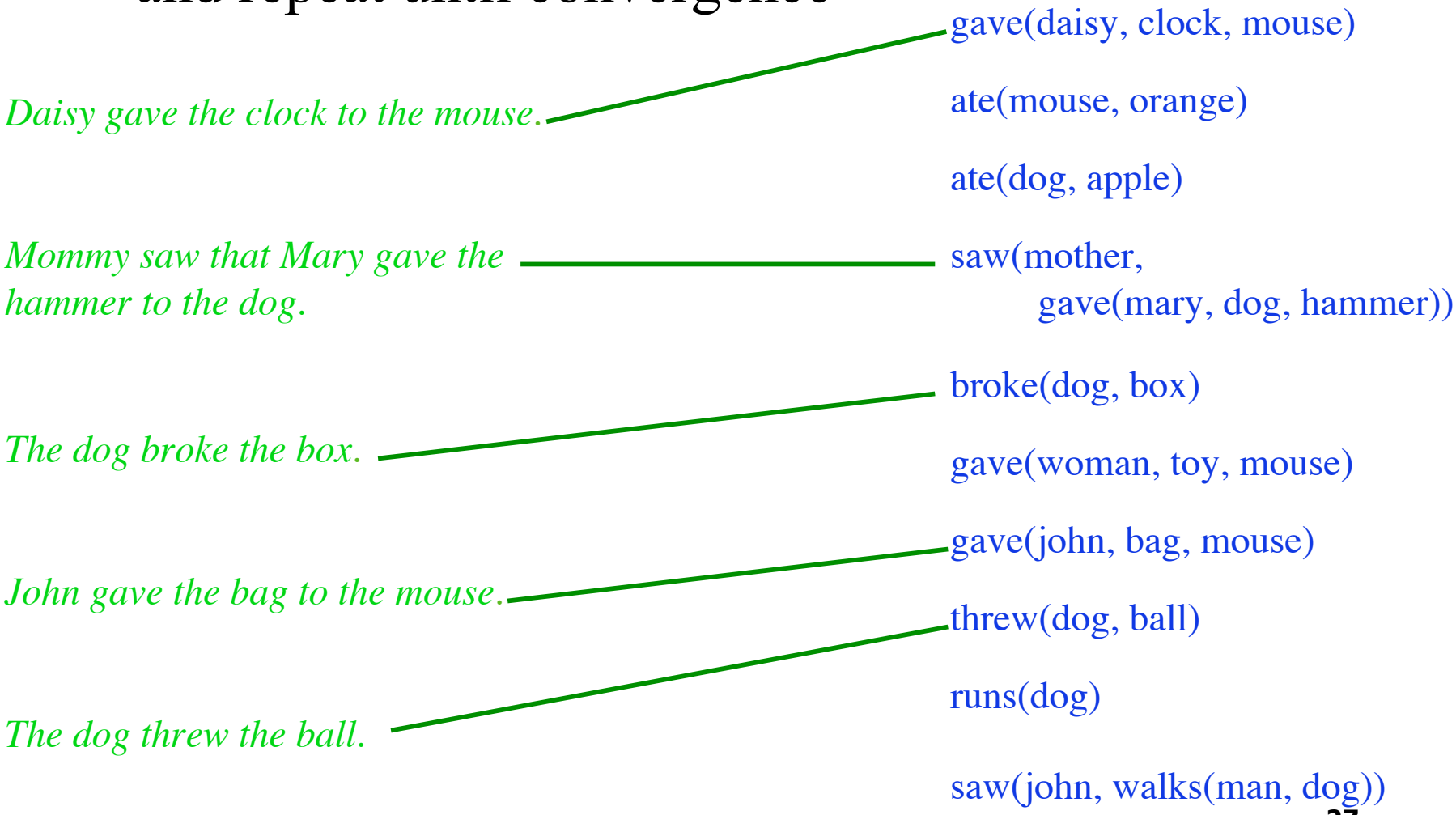
# KRISPER's Training Algorithm

4. Use *maximum weighted matching* on a bipartite graph to find the best NL-MR pairs [Munkres, 1957]



# KRISPER's Training Algorithm

5. Give the best pairs to KRISP in the next iteration, and repeat until convergence



# WASPER

- WASP with EM-like retraining to handle ambiguous training data.
- Same augmentation as added to KRISP to create KRISPER.

# KRISPER-WASP

- First iteration of EM-like training produces very noisy training data (> 50% errors).
- KRISP is better than WASP at handling noisy training data.
  - SVM prevents overfitting.
  - String kernel allows partial matching.
- But KRISP does not support language generation.
- First train KRISPER just to determine the best NL→MR matchings.
- Then train WASP on the resulting unambiguously supervised data.

# WASPER-GEN

- In KRISPER and WASPER, the correct MR for each sentence is chosen based on maximizing the confidence of semantic parsing (NL→MR).
- Instead, WASPER-GEN determines the best matching based on **generation** (MR→NL).
- Score each potential NL/MR pair by using the currently trained WASP$^{-1}$ generator.
- Compute NIST MT score between the generated sentence and the potential matching sentence.

# Strategic Generation

- Generation requires not only knowing ***how*** to say something (***tactical generation***) but also ***what*** to say (***strategic generation***).
- For automated sportscasting, one must be able to effectively choose which events to describe.

# Example of Strategic Generation

**pass ( purple7 , purple6 )**

**ballstopped**

**kick ( purple6 )**

**pass ( purple6 , purple2 )**

**ballstopped**

**kick ( purple2 )**

**pass ( purple2 , purple3 )**

**kick ( purple3 )**

**badPass ( purple3 , pink9 )**

**turnover ( purple3 , pink9 )**

# Example of Strategic Generation

**pass ( purple7 , purple6 )**

**ballstopped**

**kick ( purple6 )**

**pass ( purple6 , purple2 )**

**ballstopped**

**kick ( purple2 )**

**pass ( purple2 , purple3 )**

**kick ( purple3 )**

**badPass ( purple3 , pink9 )**

**turnover ( purple3 , pink9 )**

# Learning for Strategic Generation

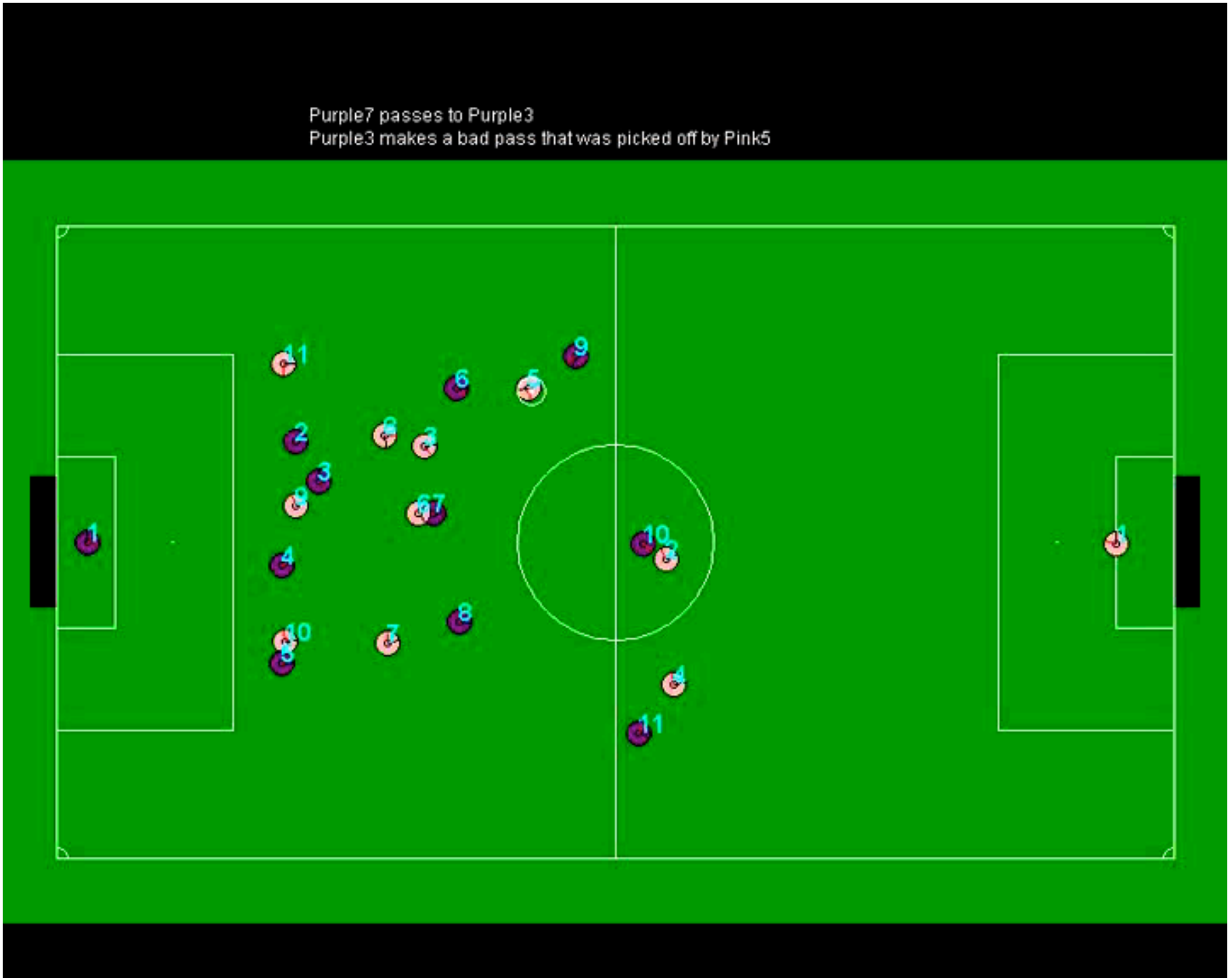
- For each event type (e.g. pass, kick) estimate the probability that it is described by the sportscaster.
- Requires NL/MR matching that indicates which events were described, but this is not provided in the ambiguous training data.
  - Use estimated matching computed by KRISPER, WASPER or WASPER-GEN.
  - Use a version of EM to determine the probability of mentioning each event type just based on strategic info.

# Iterative Generation Strategy Learning (IGSL)

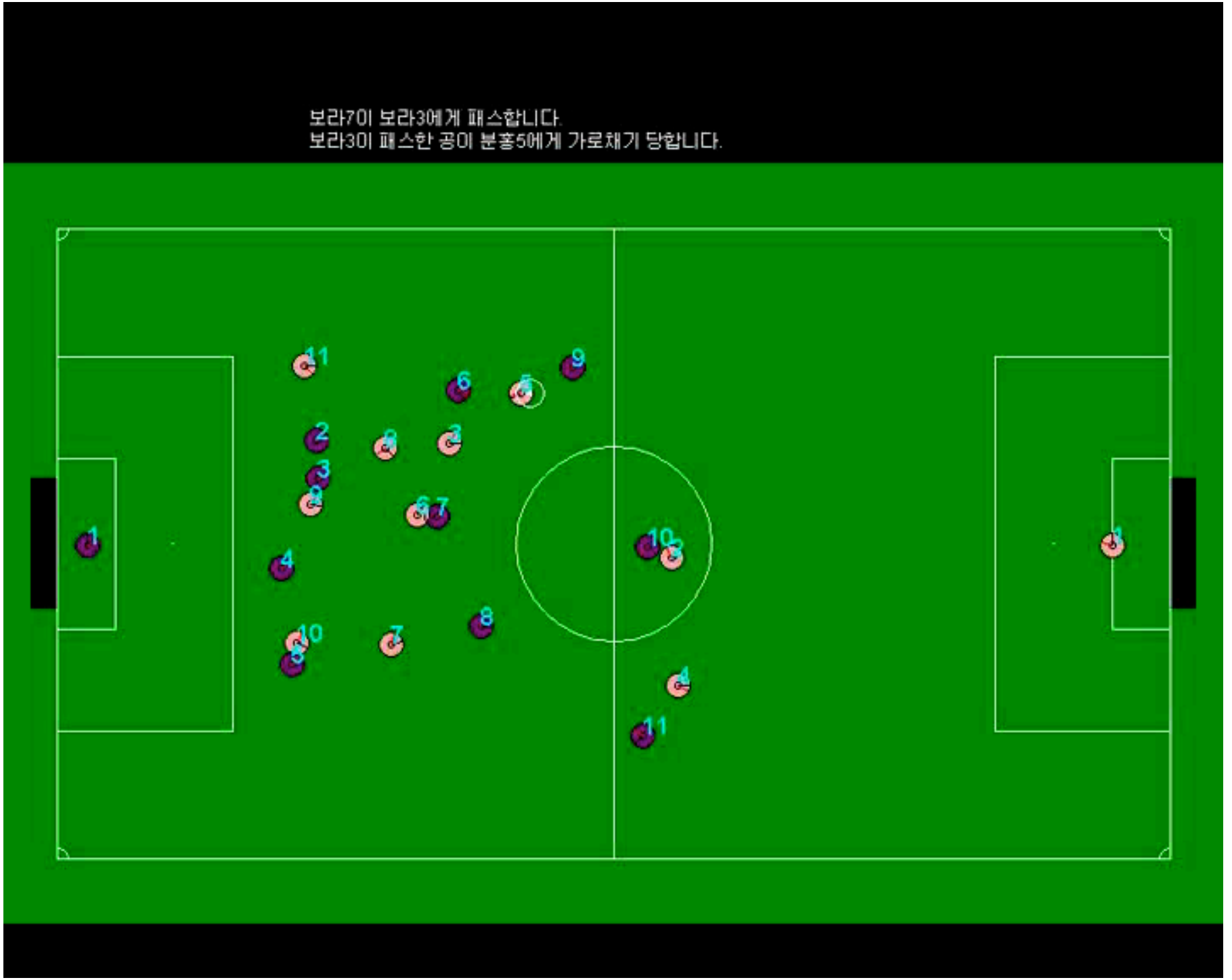
- Directly estimates the likelihood of commenting on each event type from the ambiguous training data.
- Uses self-training iterations to improve estimates (à la EM).

# Demo

- Game clip commentated using WASPER-GEN with EM-based strategic generation, since this gave the best results for generation.
- FreeTTS was used to synthesize speech from textual output.
- Also trained for Korean to illustrate language independence.


Purple7 passes to Purple3
Purple3 makes a bad pass that was picked off by Pink5

보라7이 보라3에게 패스합니다.
보라3이 패스한 공이 분홍5에게 가로채기 당합니다.



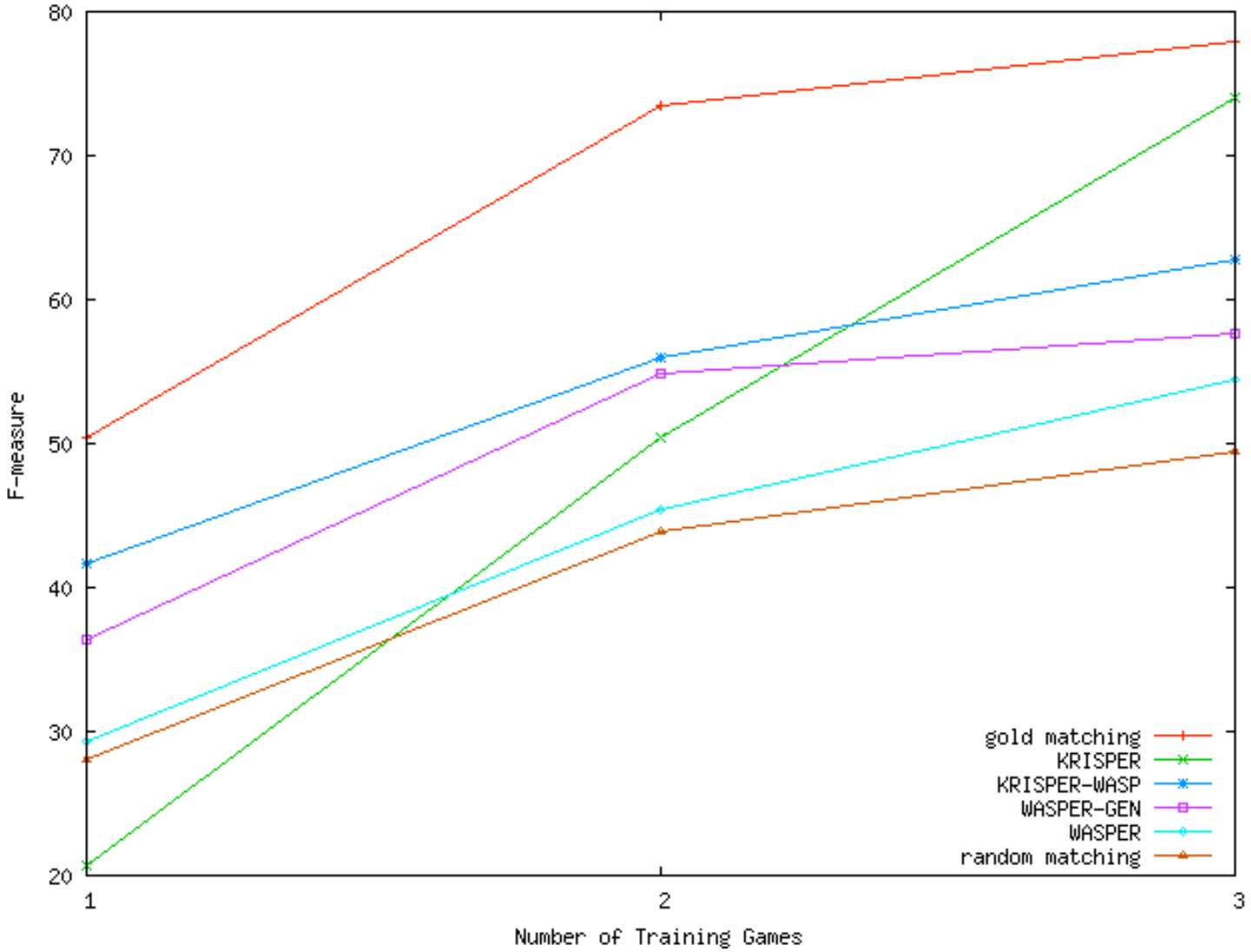
# Experimental Evaluation

- Generated learning curves by training on all combinations of 1 to 3 games and testing on all games not used for training.
- Baselines:
  - **Random Matching**: WASP trained on random choice of possible MR for each comment.
  - **Gold Matching**: WASP trained on correct matching of MR for each comment.
- Metrics:
  - **Precision**: % of system's annotations that are correct
  - **Recall**: % of gold-standard annotations correctly produced
  - **F-measure**: Harmonic mean of precision and recall

# Evaluating Semantic Parsing

- Measure how accurately learned parser maps sentences to their correct meanings in the test games.
- Use the gold-standard matches to determine the correct MR for each sentence that has one.
- Generated MR must ***exactly*** match gold-standard to count as correct.

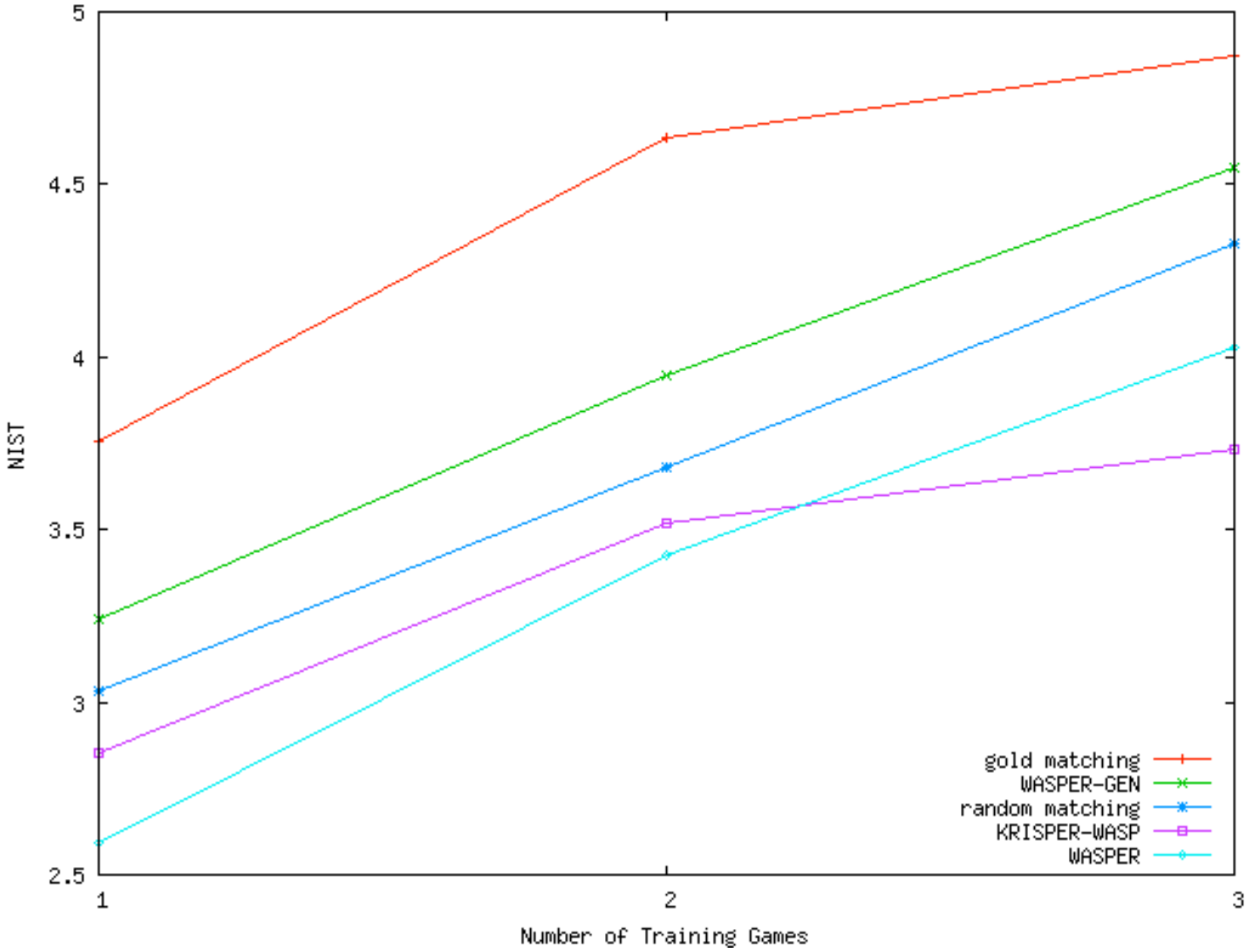
# Results on Semantic Parsing



# Evaluating Tactical Generation

- Measure how accurately NL generator produces English sentences for chosen MRs in the test games.
- Use gold-standard matches to determine the correct sentence for each MR that has one.
- Use NIST score to compare generated sentence to the one in the gold-standard.

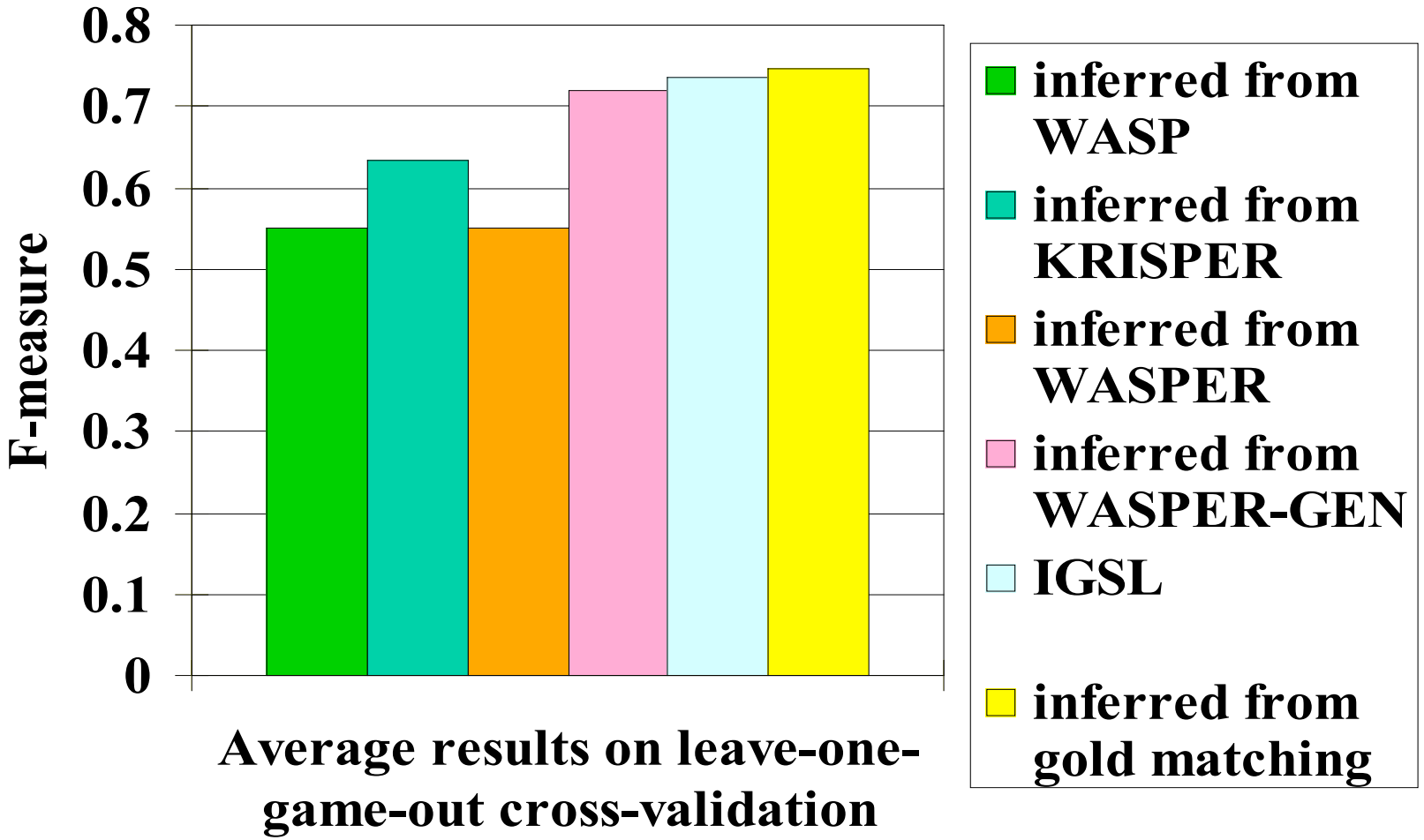
# Results on Tactical Generation



# Evaluating Strategic Generation

- In the test games, measure how accurately the system determines which perceived events to comment on.
- Compare the subset of events chosen by the system to the subset chosen by the human annotator (as given by the gold-standard matching).

# Results on Strategic Generation



# Human Evaluation (Quasi Turing Test)

- Asked 4 fluent English speakers to evaluate overall quality of sportscasts.
- Randomly picked a 2 minute segment from each of the 4 games.
- Each human judge evaluated 8 commented game clips, each of the 4 segments commented once by a human and once by the machine when tested on that game (and trained on the 3 other games).
- The 8 clips presented to each judge were shown in random counter-balanced order.
- Judges were not told which ones were human or machine generated.

# Human Evaluation Metrics

| Score | English Fluency | Semantic Correctness | Sportscasting Ability |
|---|---|---|---|
| 5 | Flawless | Always | Excellent |
| 4 | Good | Usually | Good |
| 3 | Non-native | Sometimes | Average |
| 2 | Disfluent | Rarely | Bad |
| 1 | Gibberish | Never | Terrible |

# Results on Human Evaluation

| Commentator | English Fluency | Semantic Correctness | Sportscasting Ability |
|---|---|---|---|
| Human | 3.94 | 4.25 | 3.63 |
| Machine | 3.44 | 3.56 | 2.94 |
| Difference | –0.5 | –0.69 | –0.69 |

# Co-Training with Visual and Textual Views (Gupta, Kim, Grauman & Mooney, ECML-08)

# Semi-Supervised Multi-Modal Image Classification

- Use both images or videos and their textual captions for classification.
- Use semi-supervised learning to exploit unlabeled training data in addition to labeled training data.
- How?: ***Co-training*** (Blum and Mitchell, 1998) using visual and textual views.
- Illustrates both language supervising vision and vision supervising language.

# Sample Classified Captioned Images

**Desert**




Cultivating farming at Nabataean Ruins of the Ancient Avdat




Bedouin Leads His Donkey That Carries Load Of Straw

**Trees**




Ibex Eating In The Nature




Entrance To Mikveh Israel Agricultural School

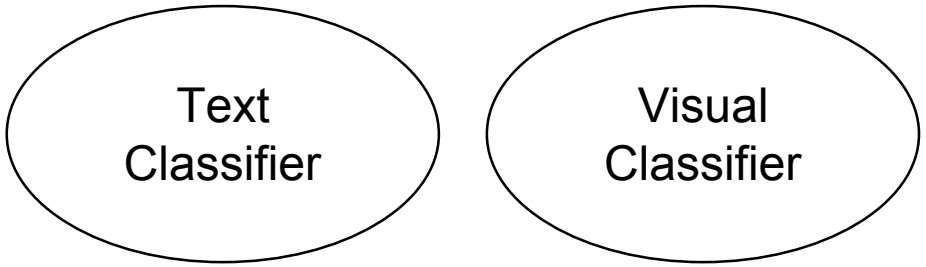
# Co-training

- Semi-supervised learning paradigm that exploits two mutually independent and sufficient views

- Features of dataset can be divided into two sets:
  - The instance space: $X = X_1 \times X_2$
  - Each example: $x = (x_1, x_2)$

- Proven to be effective in several domains
  - Web page classification (content and hyperlink)
  - E-mail classification (header and body)

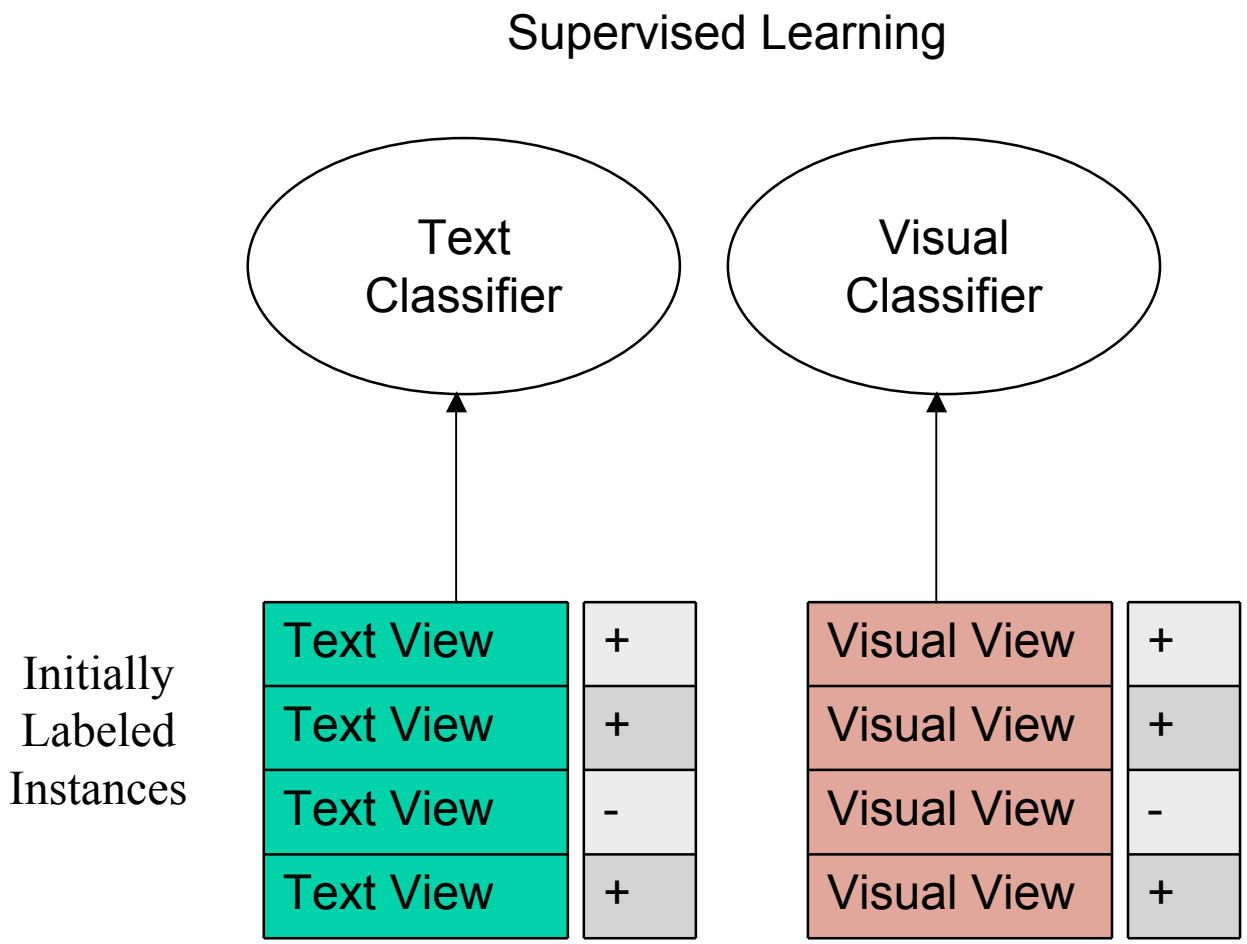
# Co-training



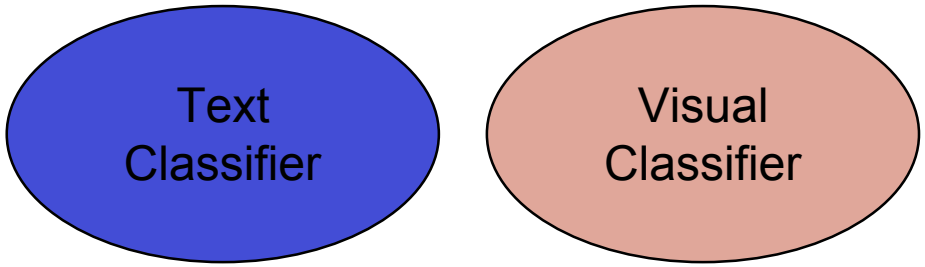

Initially Labeled Instances

| | | |
|---|---|---|
| Text View | Visual View | + |
| Text View | Visual View | + |
| Text View | Visual View | - |
| Text View | Visual View | + |

# Co-training


Supervised Learning
Text Classifier
Visual Classifier
Initially Labeled Instances
Text View
Text View
Text View
Text View
+
+
-
+
Visual View
Visual View
Visual View
Visual View
+
+
-
+

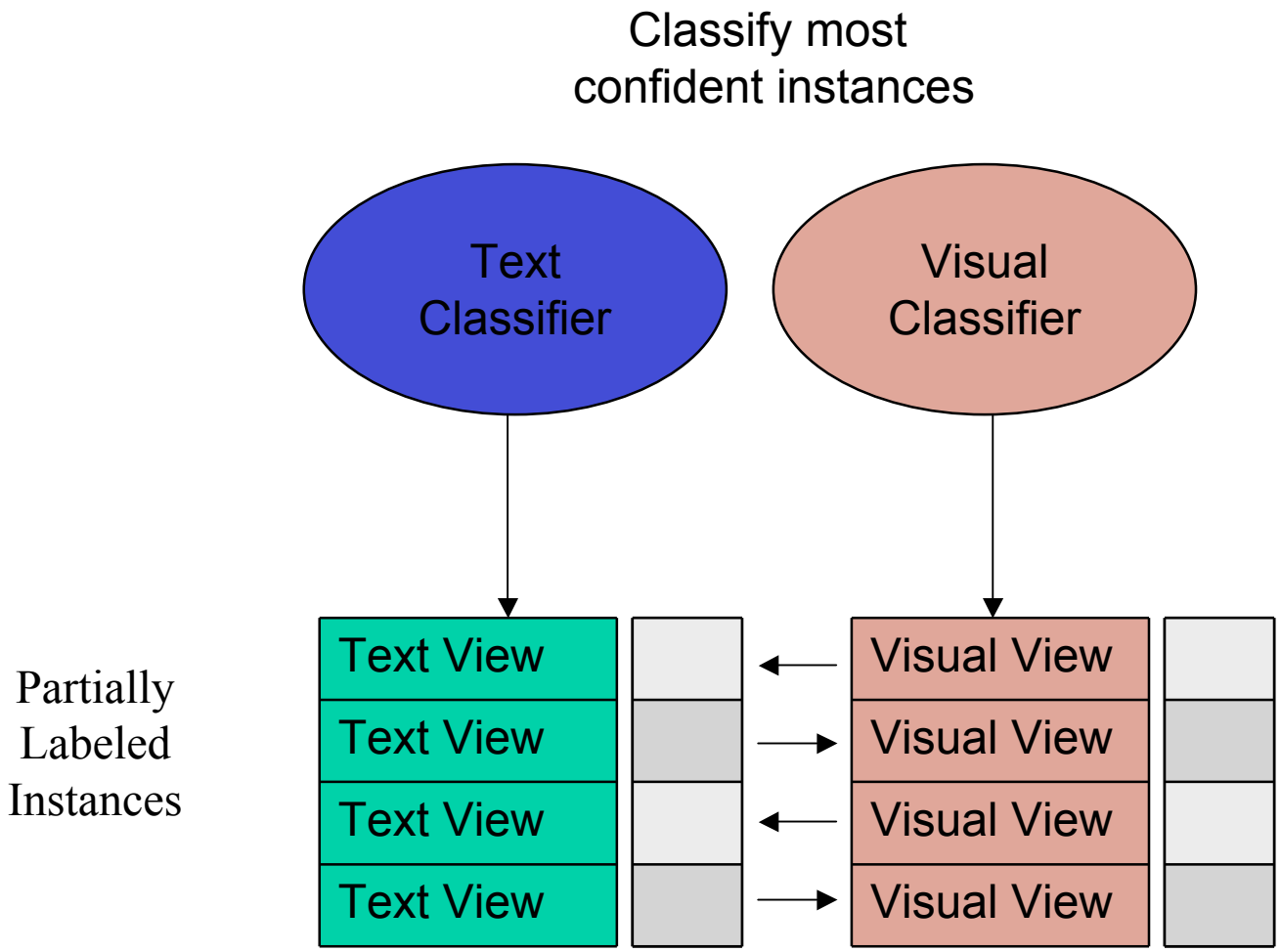
# Co-training



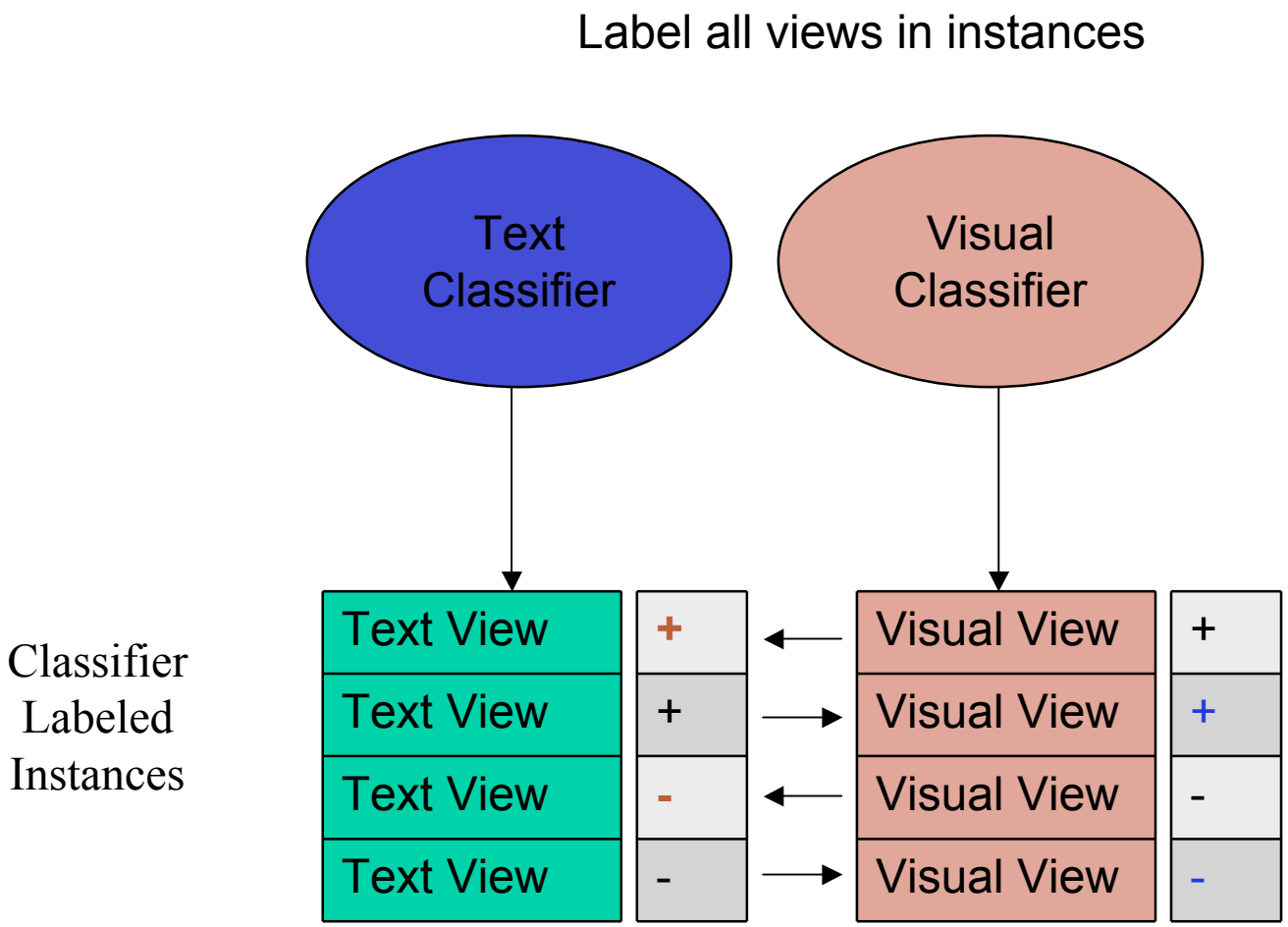

Unlabeled Instances

| | | |
|---|---|---|
| Text View | Visual View | |
| Text View | Visual View | |
| Text View | Visual View | |
| Text View | Visual View | |

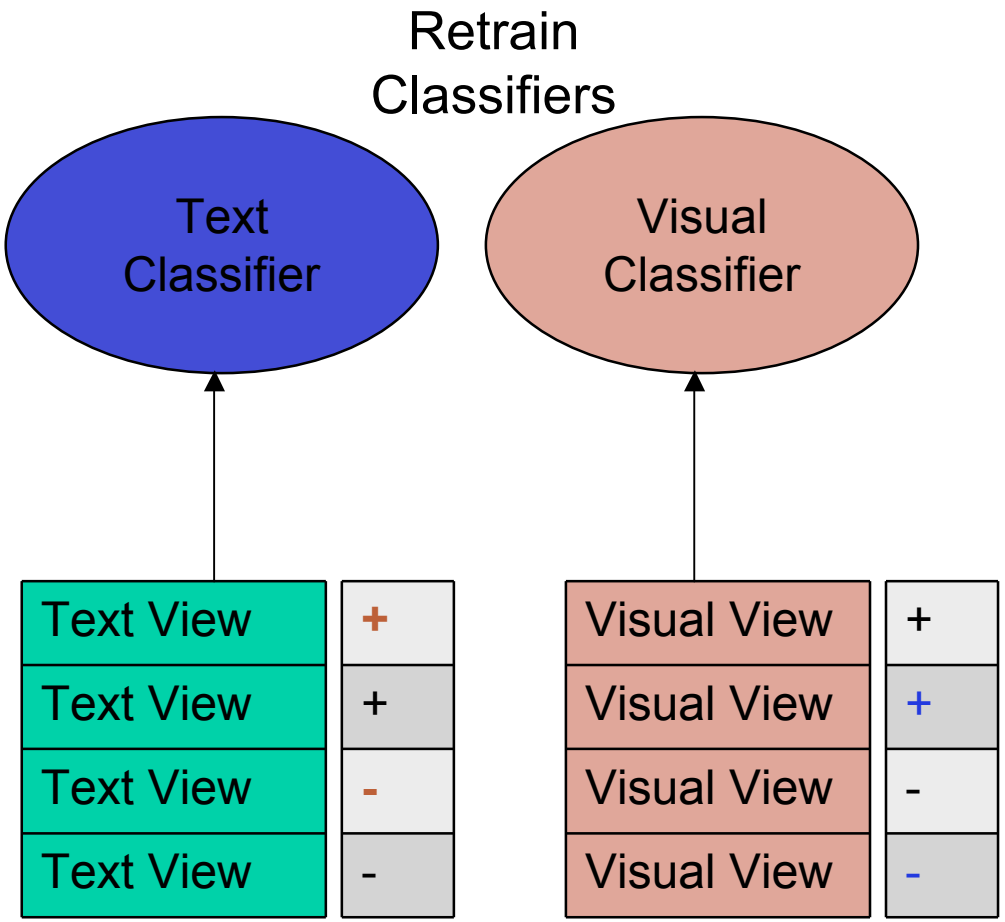
# Co-training


Classify most
confident instances
Text
Classifier
Visual
Classifier
Partially
Labeled
Instances
Text View
Text View
Text View
Text View
Visual View
Visual View
Visual View
Visual View

# Co-training

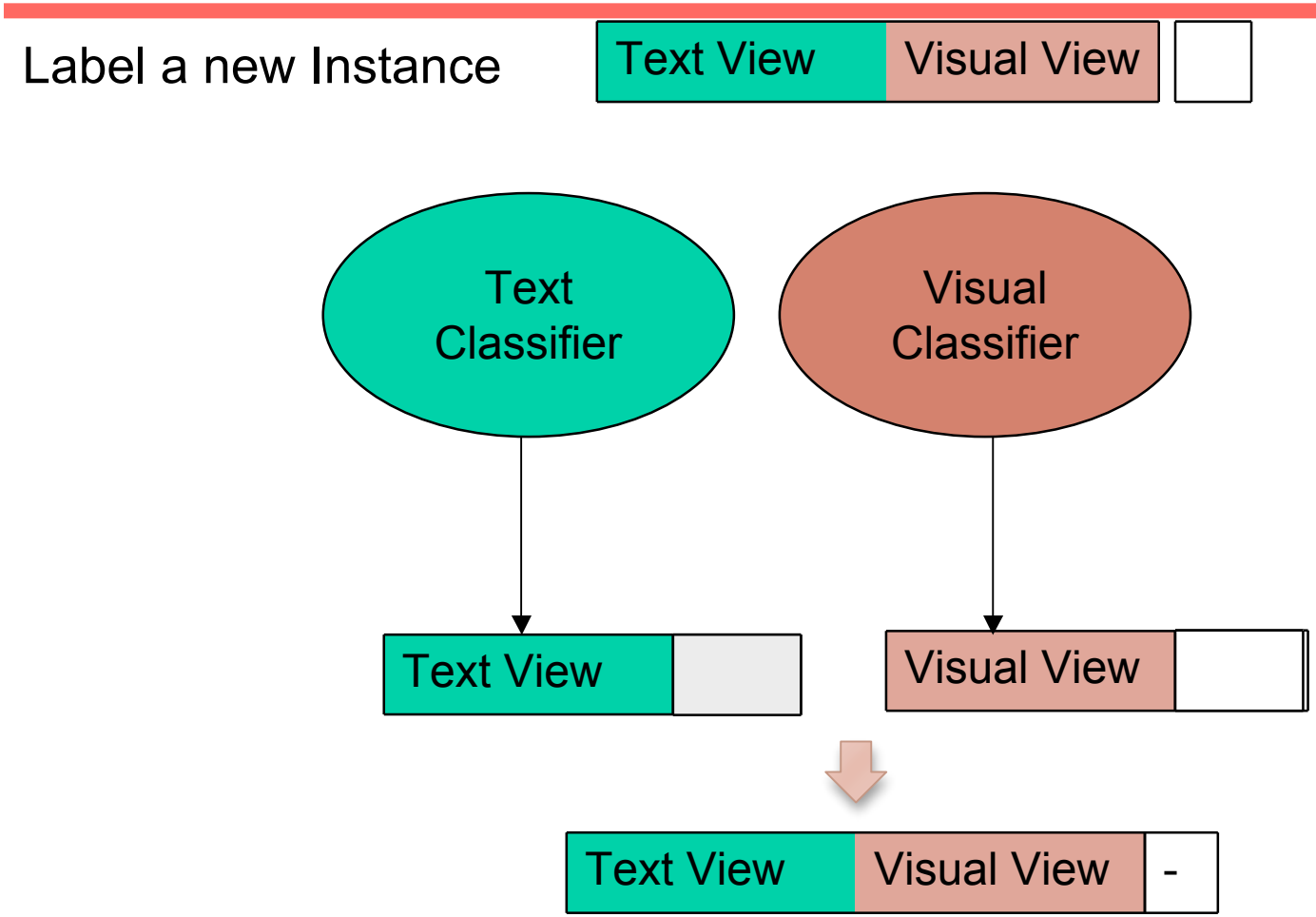

Label all views in instances
Text Classifier
Visual Classifier
Classifier Labeled Instances
Text View
Text View
Text View
Text View
+
+
-
-
Visual View
Visual View
Visual View
Visual View
+
+
-
-

# Co-training


Retrain
Classifiers
Text
Classifier
Visual
Classifier
Text View
Text View
Text View
Text View
+
+
-
-
Visual View
Visual View
Visual View
Visual View
+
+
-
-

# Co-training

Label a new Instance

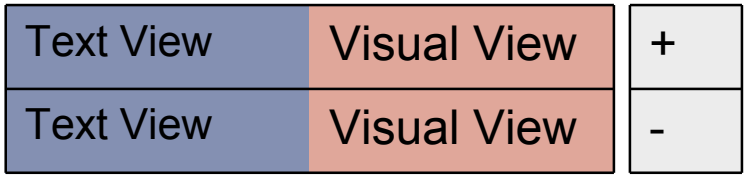


# Baseline - Individual Views

- **Image/Video View** : Only image/video features are used
- **Text View** : Only textual features are used

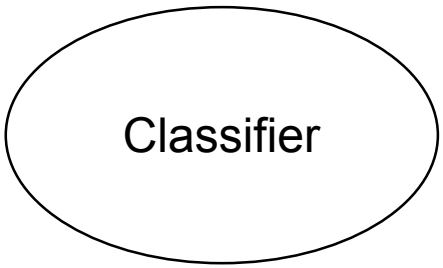
# Baseline - Early Fusion

- Concatenate visual and textual features



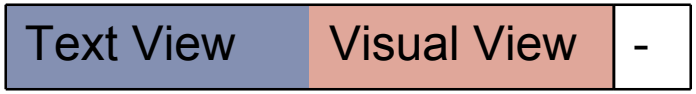



Training

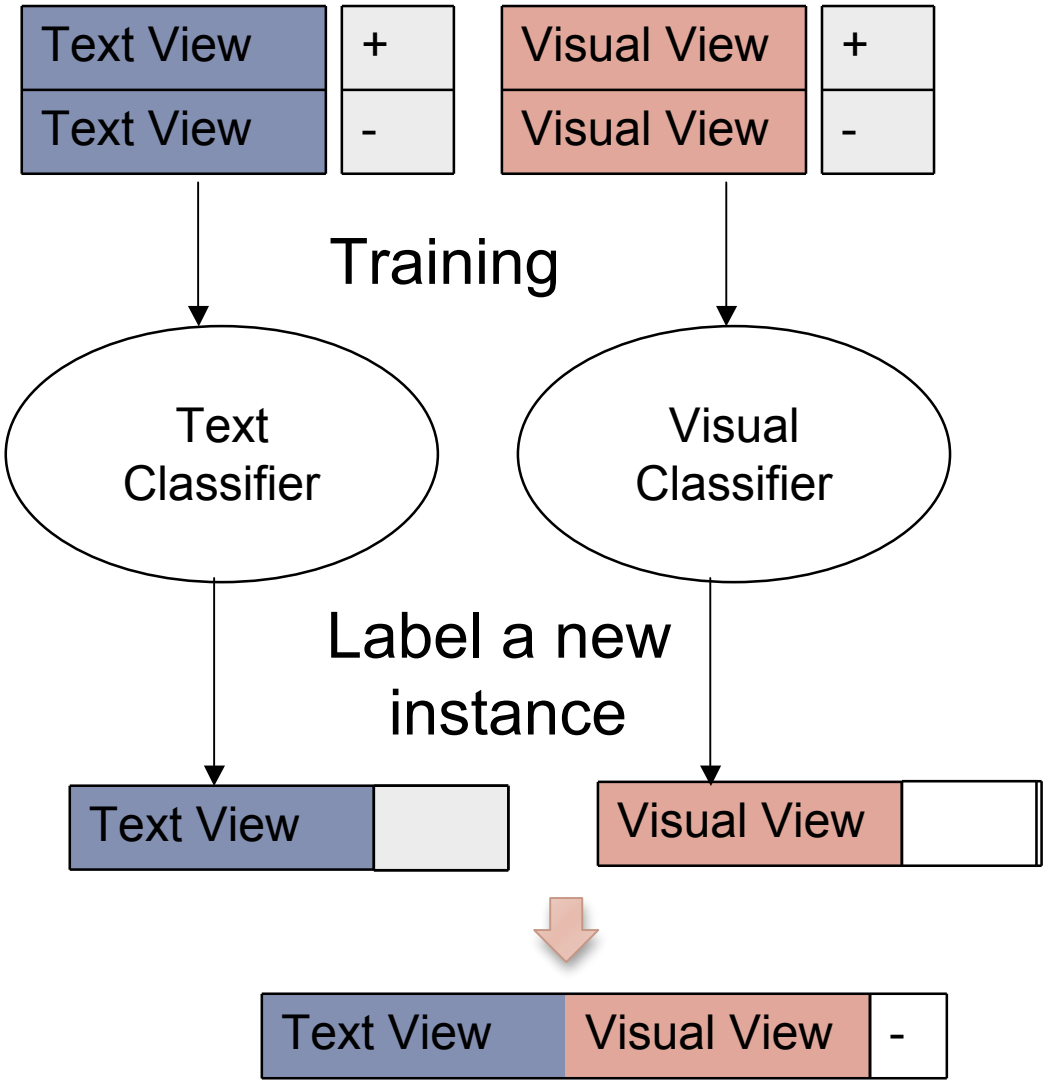





Testing



# Baseline - Late Fusion



# Image Dataset

- Our captioned image data is taken from (Bekkerman & Jeon CVPR '07, www.israelimages.com)
- Consists of images with short text captions.
- Used two classes, Desert and Trees.
- A total of 362 instances.

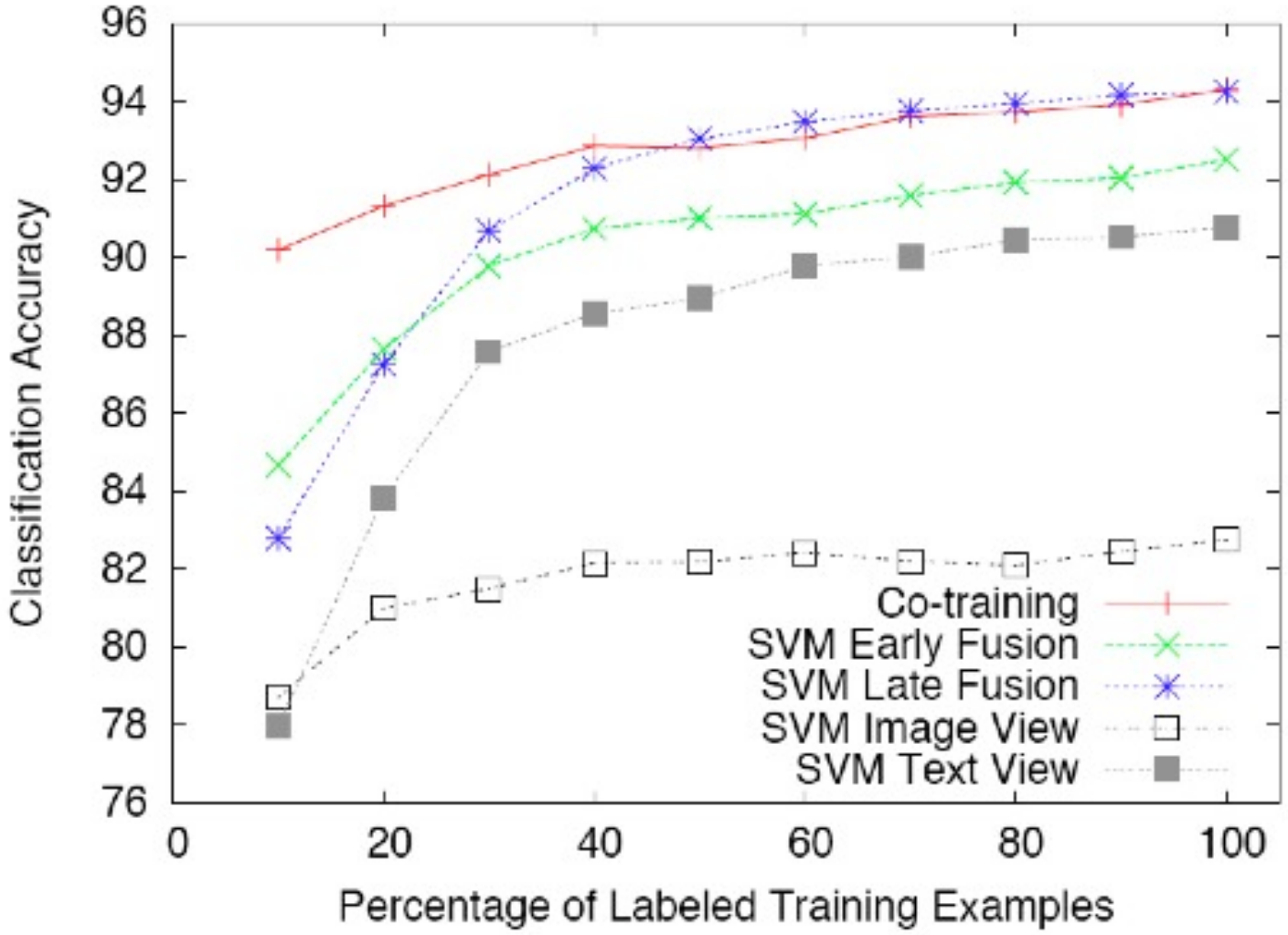
# Text and Visual Features

- **Text view**: standard bag of words.
- **Image view**: standard bag of visual words that capture texture and color information.

# Experimental Methodology

- Test set is disjoint from **both** labeled and unlabeled training set.
- For plotting learning curves, vary the percentage of training examples labeled, rest used as unlabeled data for co-training.
- SVM with RBF kernel is used as base classifier for both visual and text classifiers.
- All experiments are evaluated with 10 iterations of 10-fold cross-validation.

# Learning Curves for Israel Images



# Using Closed Captions to Supervise Activity Recognition in Videos (Gupta & Mooney, VCL-09)

# Activity Recognition in Video

- Recognizing activities in video generally uses supervised learning trained on human-labeled video clips.
- Linguistic information in closed captions (CCs) can be used as "weak supervision" for training activity recognizers.
- Automatically trained activity recognizers can be used to improve precision of video retrieval.

# Sample Soccer Videos

## Kick




"I do not think there is any real intent, just trying to make sure he gets his body across, but it was a free **kick** ."




"Lovely **kick**."




"Goal **kick**."

## Save




"Good **save** as well."




"I think brown made a wonderful fingertip **save** there."




"And it is a really chopped **save**"

# Throw



"If you are defending a lead, your **throw** back takes it that far up the pitch and gets a **throw**-in."



"Another shot for a **throw**."



"And Carlos Tevez has won the **throw**."

# Touch



"All it needed was a **touch**."



"When they are going to pass it in the back, it is a really pure **touch**."



"Look at that, Henry, again, he had time on the ball to take another **touch** and prepare that ball properly."

# Using Video Closed-Captions

- CCs contains both relevant and irrelevant information:

  *"Beautiful pull-back."* ***relevant***

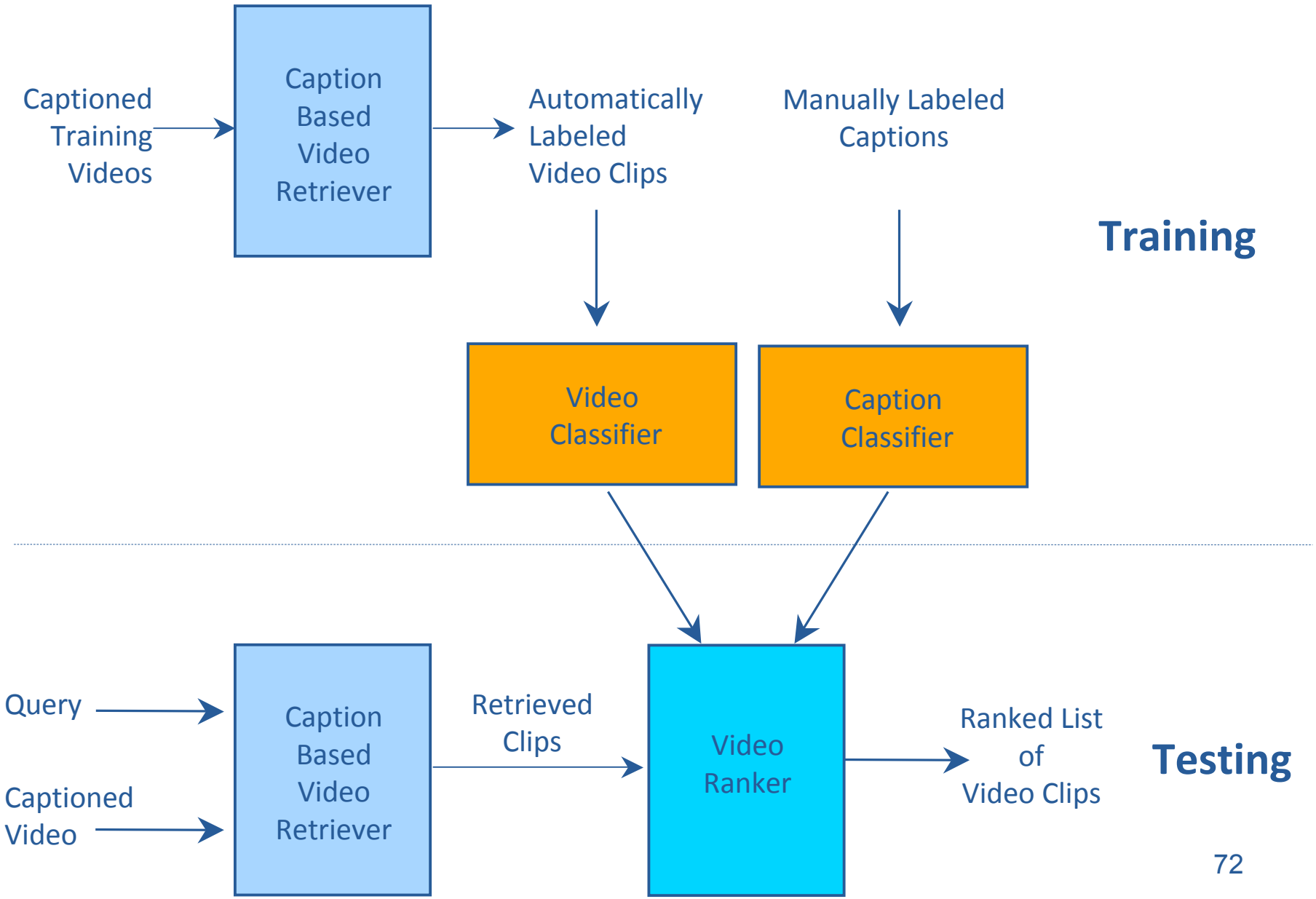
  *"They scored in the last kick of the game against the Czech Republic."* ***irrelevant***

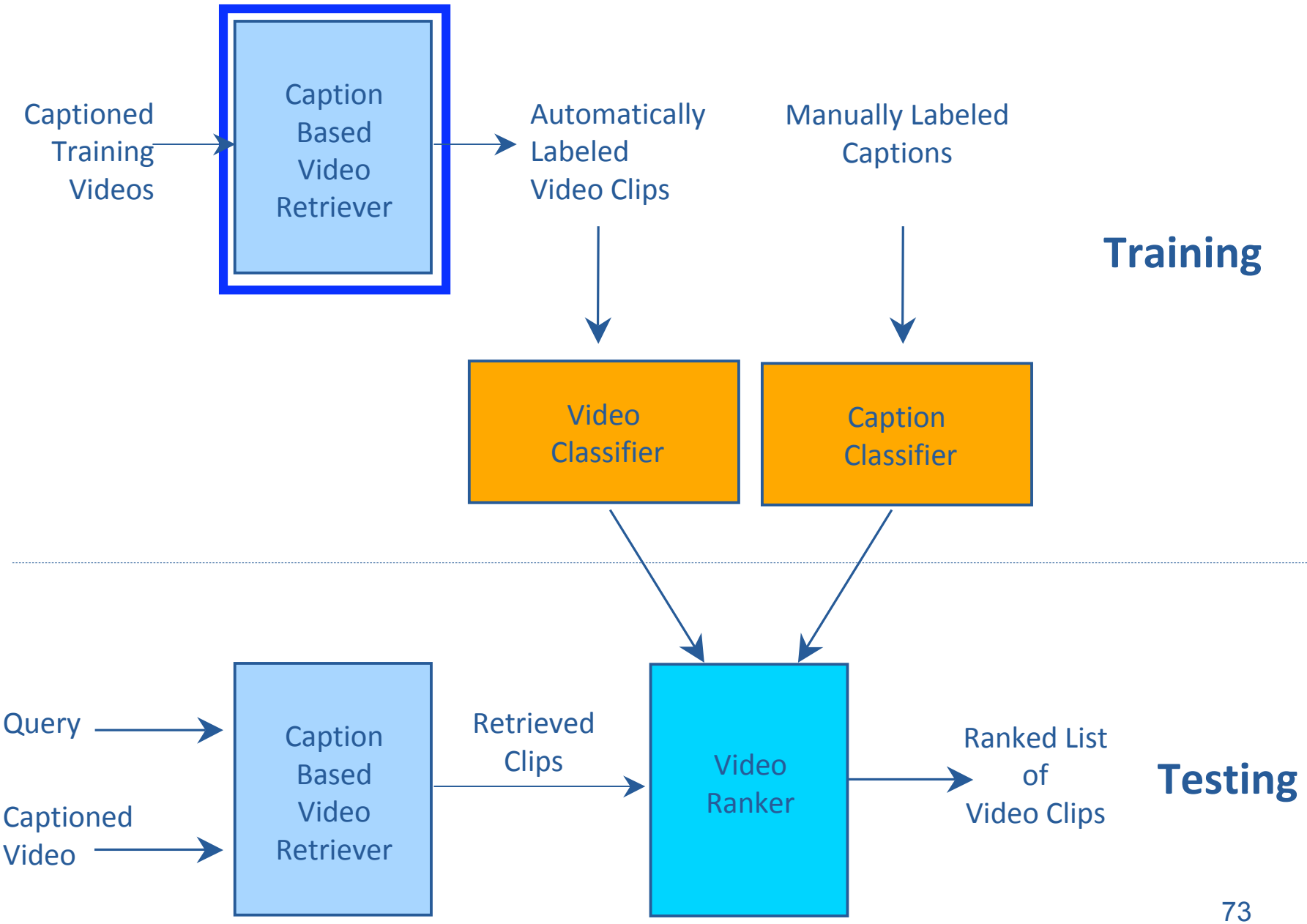
  *"That is a fairly good tackle."* ***relevant***

  *"Turkey can be well-pleased with the way they started."* ***irrelevant***

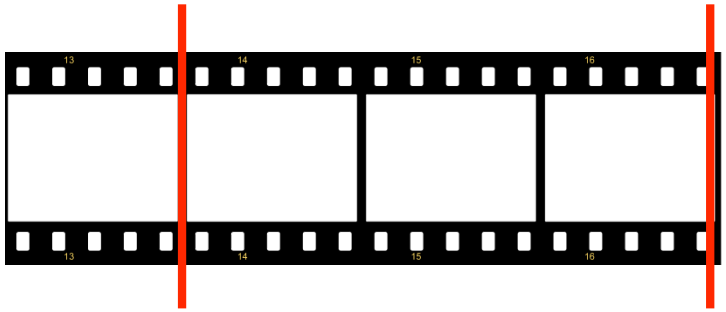
- Use a novel caption classifier to ***rank*** the retrieved video clips by relevance.

# SYSTEM OVERVIEW

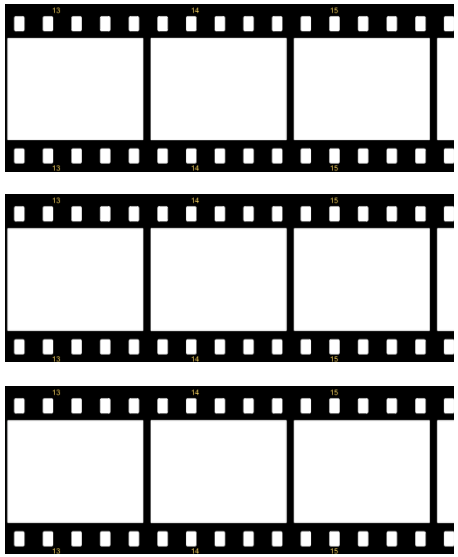



Captioned Training Videos
Caption Based Video Retriever
Automatically Labeled Video Clips
Manually Labeled Captions
Training
Video Classifier
Caption Classifier
Query
Captioned Video
Caption Based Video Retriever
Retrieved Clips
Video Ranker
Ranked List of Video Clips
Testing

# Retrieving and Labeling Data



…What a nice **kick**!…

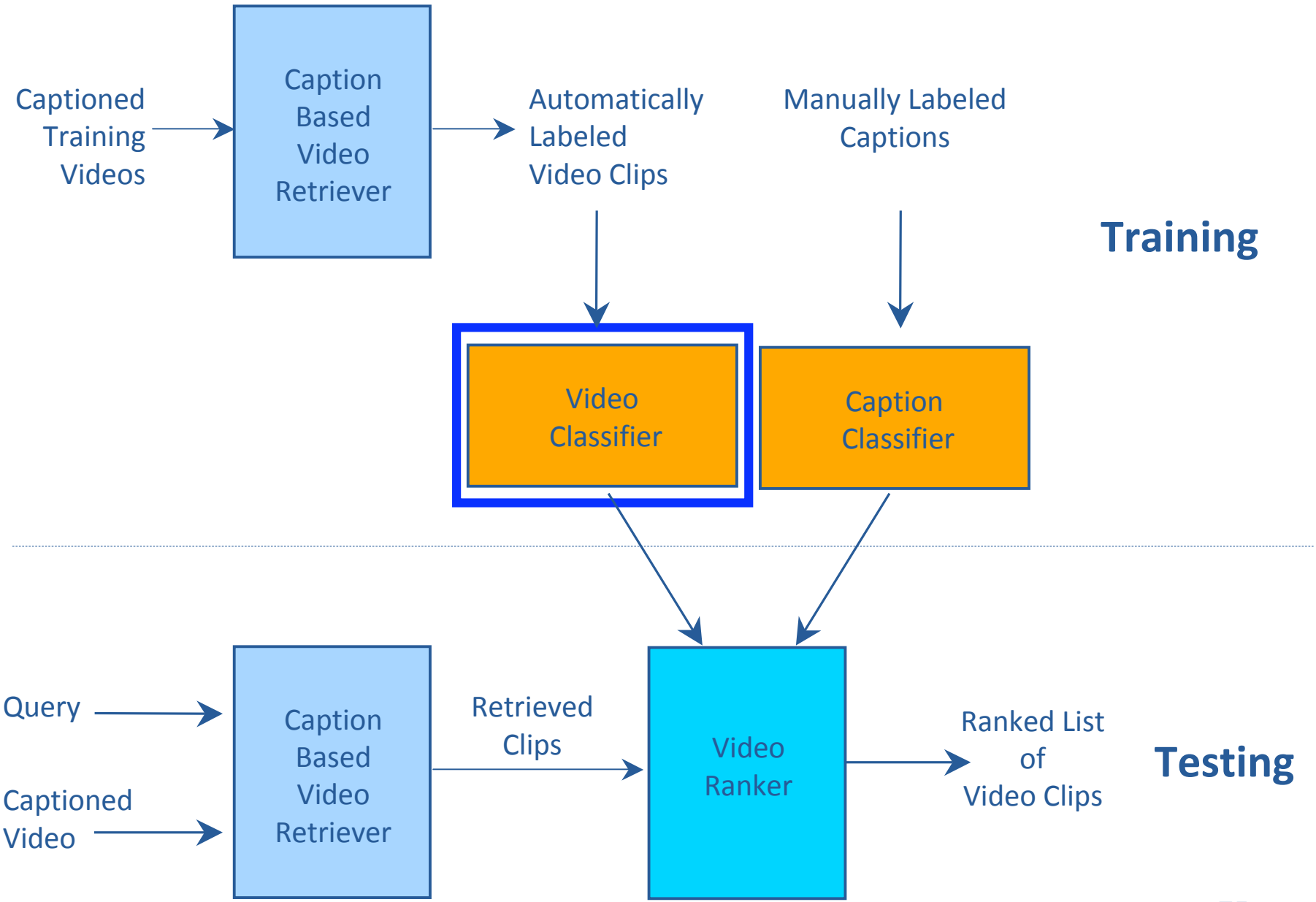


kick

save

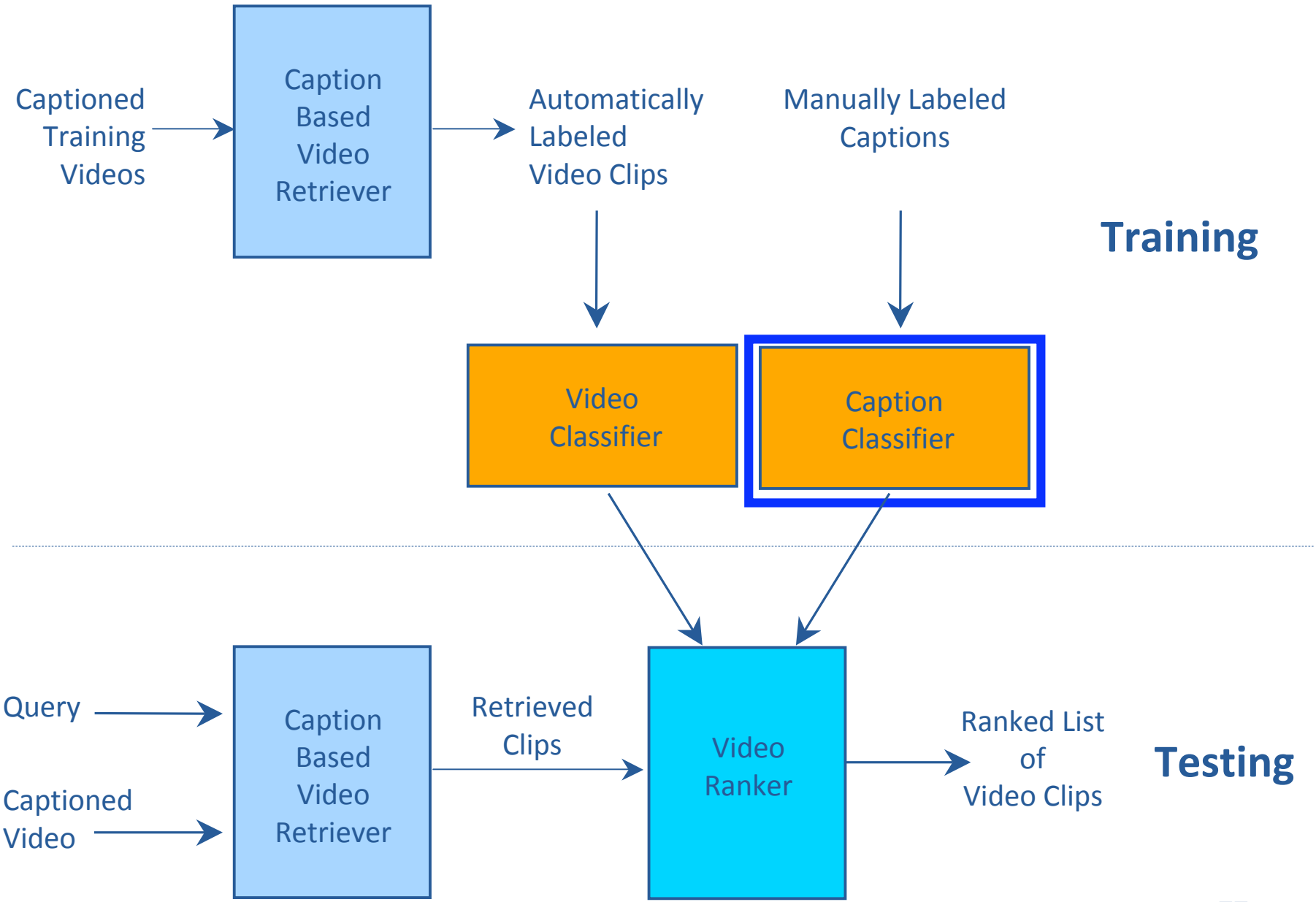
touch

- Identify all closed caption sentences that contain exactly one of the set of activity keywords
  - kick, save, throw, touch
- Extract clips of 8 sec around the corresponding time
- Label the clips with corresponding classes


Captioned Training Videos
Caption Based Video Retriever
Automatically Labeled Video Clips
Manually Labeled Captions
Training
Video Classifier
Caption Classifier
Query
Captioned Video
Caption Based Video Retriever
Retrieved Clips
Video Ranker
Ranked List of Video Clips
Testing

# Video Classifier

- Extract visual features from clips.
  - Histogram of oriented gradients and optical flow in space-time volume (Laptev *et al.*, ICCV 07; CVPR 08)
  - Represent as 'bag of visual words'
- Use automatically labeled video clips to train activity classifier.
- Use DECORATE (Melville and Mooney, IJCAI 03 )
  - An ensemble based classifier
  - Works well with **noisy** and limited training data


Captioned Training Videos
Caption Based Video Retriever
Automatically Labeled Video Clips
Manually Labeled Captions
Training
Video Classifier
Caption Classifier
Query
Captioned Video
Caption Based Video Retriever
Retrieved Clips
Video Ranker
Ranked List of Video Clips
Testing

# Caption Classifier

- Sportscasters talk about both events on the field as well as other information
  - 69% of the captions in our dataset are 'irrelevant' to the current events
- Classifies relevant vs. irrelevant captions
  - Independent of the query classes
- Use SVM string classifier
  - Uses a subsequence kernel that measures how many subsequences are shared by two strings (Lodhi *et al*. 02, Bunescu and Mooney 05)
  - More accurate than a "bag of words" classifier since it takes word order into account.

# Retrieving and Ranking Videos

- Videos retrieved using captions, same way as before.
- Two ways of ranking:
  - Probabilities given by video classifier (VIDEO)
  - Probabilities given by caption classifier (CAPTION)
- Aggregating the rankings
  - Weighted late fusion of rankings from VIDEO and CAPTION

$$P(label \mid clip\text{-}with\text{-}caption) = \alpha P(label \mid clip) + (1-\alpha) P(relevant \mid clip\text{-}caption)$$

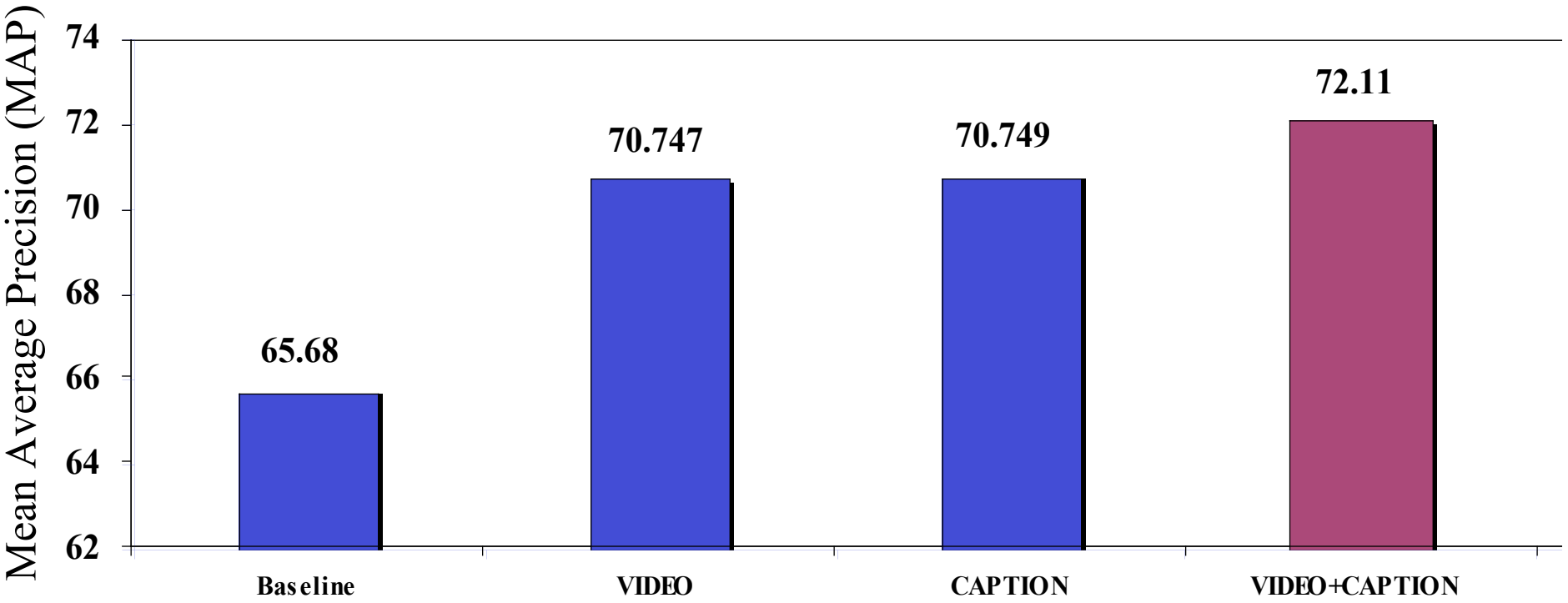
# Experiment

- Dataset
  - 23 soccer games recorded from TV broadcast
  - Avg. length: 1 hr 50 min
  - Avg. number of captions: 1,246
  - Caption Classifier
    - Trained on hand labeled 4 separate games
- Metric: MAP score: Mean Averaged Precision
- Methodology: Leave one-game-out cross-validation
- Baseline: ranking clips randomly

# Dataset Statistics

| Query | # Total | # Correct | % Noise |
|---|---|---|---|
| Kick | 303 | 120 | 60.39 |
| Save | 80 | 47 | 41.25 |
| Throw | 58 | 26 | 55.17 |
| Touch | 183 | 122 | 33.33 |

# Retrieval Results



# Future Work

- Use real (not simulated) visual context to supervise language learning.
- Use more sophisticated linguistic analysis to supervise visual learning.

# Conclusions

- Current language and visual learning uses expensive, unrealistic training data.
- Naturally occurring perceptual context can be used to supervise language learning:
  - Learning to sportscast simulated Robocup games.
- Naturally occurring linguistic context can be used to supervise learning for computer vision:
  - Using multi-modal co-training to improve classification of captioned images and videos.
  - Using closed-captions to automatically train activity recognizers and improve video retrieval.

# Questions?

Relevant Papers at:

**http://www.cs.utexas.edu/users/ml/publication/clamp.html**