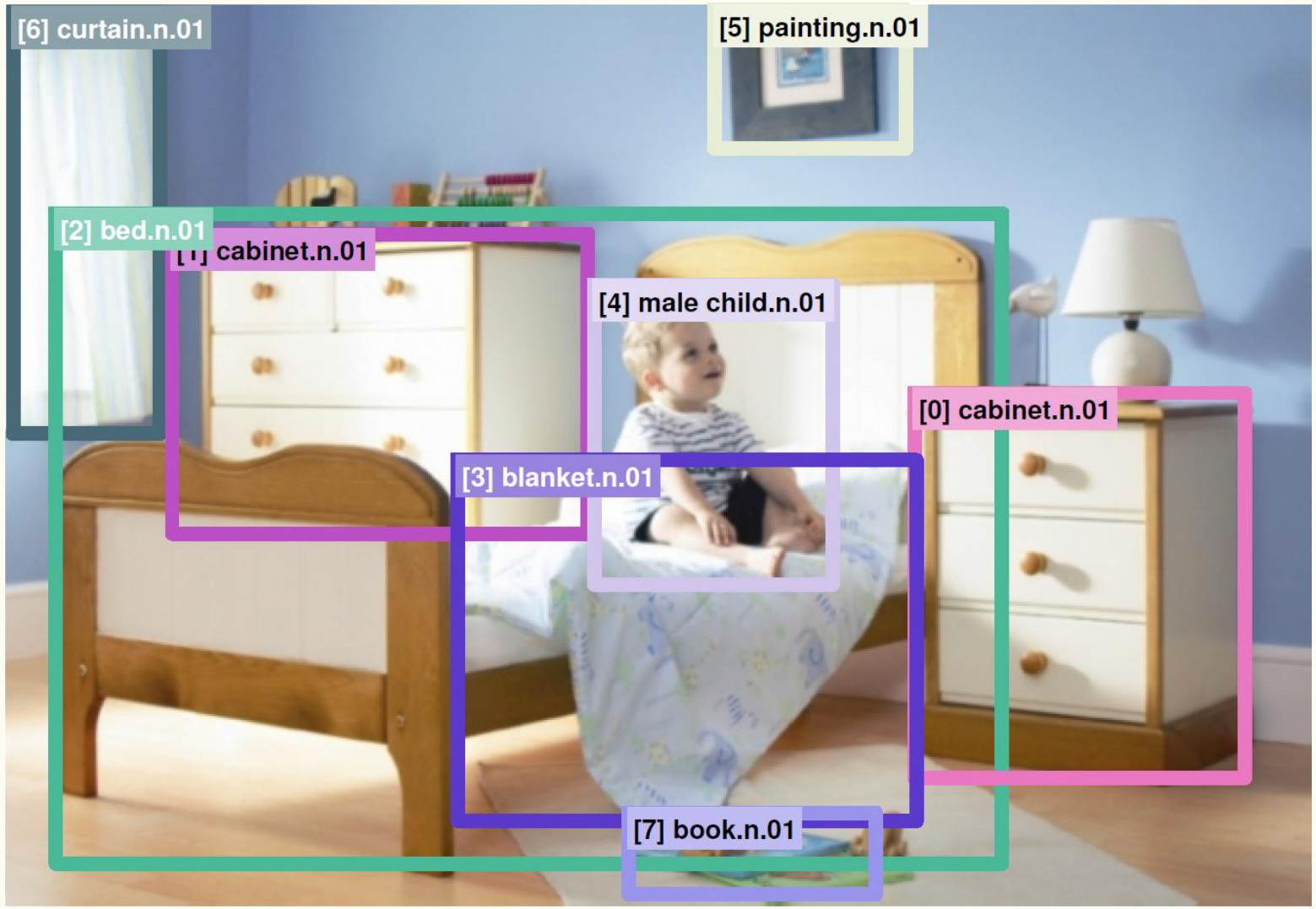
Generating Image Descriptions with Gold Standard Visual Inputs: Motivation, Evaluation and Baselines

Josiah Wang & Robert Gaizauskas

The University Of Sheffield.

Work presented as part of the Visual Sense (ViSen) consortium, funded by the chist-era D2K programme

Task: Generate a description...



... given labelled bounding boxes

This Presentation

- ImageCLEF2015 image description generation task
- Fine-grained evaluation metric for content selection
- Introduce and evaluate baselines

Motivation

- Why gold standard visual input?
  - 'Detect and generate' methods – sensitive to noisy vision input
  - Assume 'perfect' input: Can we generate? How?
- Why fine-grained evaluation?
  - Human judgments are expensive, not scalable
  - Existing measures (BLEU, ROUGE, METEOR, CIDEr) are global
  - Better to evaluate tasks/phases of NLG pipeline explicitly: referring expression, verb/event/attribute/spatial relation expression, content ordering, **content selection**

ImageCLEF

- Evaluation campaign since 2003
- Benchmark automatic image annotation & indexing for wide range of source images
- Part of CLEF initiative
  - Conference and Labs of the Evaluation Forum
  - Evaluation of information access systems since 2000

ImageCLEF 2015

- Four main tasks in 2015
  - **Scalable image annotation**
  - Medical classification
  - Medical clustering
  - Liver CT annotation

ImageCLEF 2015: Scalable Image Annotation

- Training set: 500k (image, webpage) pairs
- Subtask 1: Image annotation + localisation
  - For each 500k image, annotate + localise with 251 concepts
- Subtask 2: Image description generation
  - Noisy track: Generate descriptions for all 500k images
  - **Clean track: Generate descriptions for 450 test images, given labelled bounding boxes** ⟵ **this talk**

ImageCLEF 2015: Subtask 2 (Clean Track)

- Validation set (500 images from 500k)
  - Minimum 5 descriptions per image
    - Mean 9.5: Median: 8, Max: 51
  - Labelled bounding box annotations for 251 WordNet synsets
  - Correspondence annotation between text and bounding boxes

A [woman]^{2} in a white [dress]^{0} and gold [boots]^{5} leaning on a [car]^{3} .

A [woman]^{2} poses along a [car]^{3} .

[Woman]^{2} dressed in white with gold [boots]^{5} poses next to a police [car]^{3} .

A [woman]^{2} dressed in white leans against a white [car]^{3} .

A [woman]^{2} is leaning against a [car]^{3} .

A [woman]^{2} with gold [boots]^{5} leans against an Indy pace [car]^{3} .

A blonde [woman]^{2} wearing gold shiny [boots]^{5}, a white [top]^{0} and short white skirt is leaning on a [car]^{3} .

Content Selection Evaluation Metric

bbox instances referenced in gold standard

bbox instances referenced in generated sentence

$$P^I = \frac{1}{M} \sum_{m=1}^{M} \frac{|\, G_m^I \cap S^I \,|}{|S^I|} \qquad R^I = \frac{1}{M} \sum_{m=1}^{M} \frac{|\, G_m^I \cap S^I \,|}{|G_m^I|} \qquad F^I = 2 \times \frac{P^I \times R^I}{P^I + R^I}$$

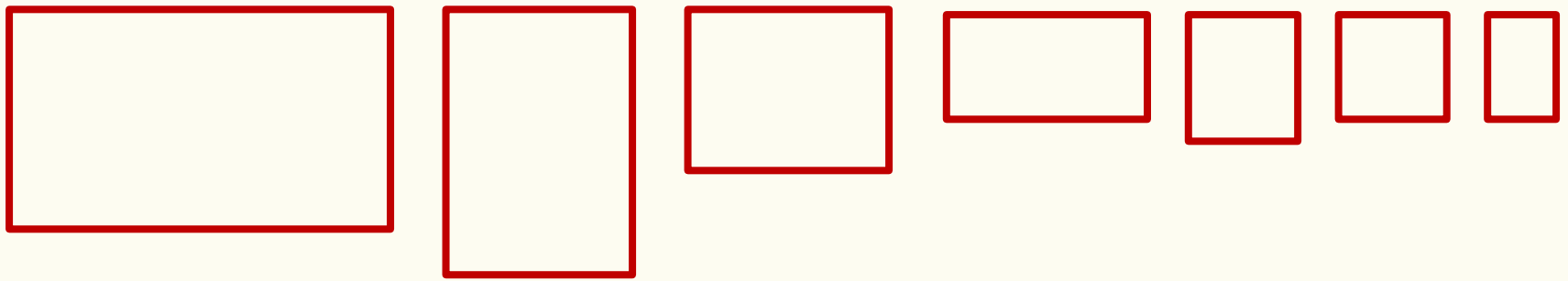
gold standard descriptions for image *I*

- Final score: Average over all test images

Generating Descriptions: Baselines

- Visual Cues
  - Bounding box size (choose biggest first)
  - Bounding box position (choose most central object first)

- Threshold selected number of bounding boxes, k



Generating Descriptions: Baselines

- Textual Priors (from validation set)
  - Unigram
  - Bigram

A [woman] in a white [dress] and gold [boots] leaning on a [car] .

<start> → woman

woman → dress

dress → boots

boots → car

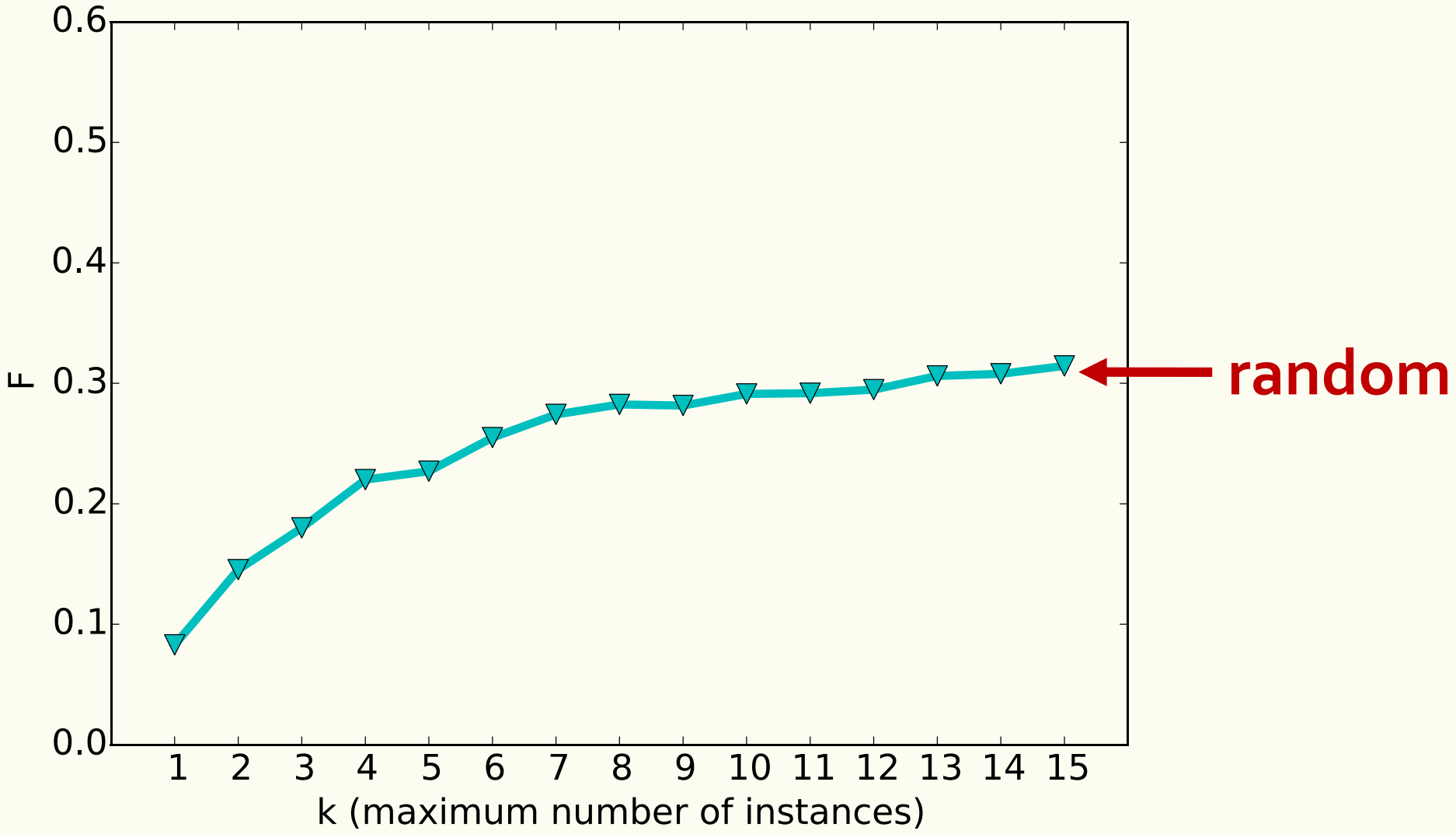
car → <end>

Generating Descriptions: Baselines

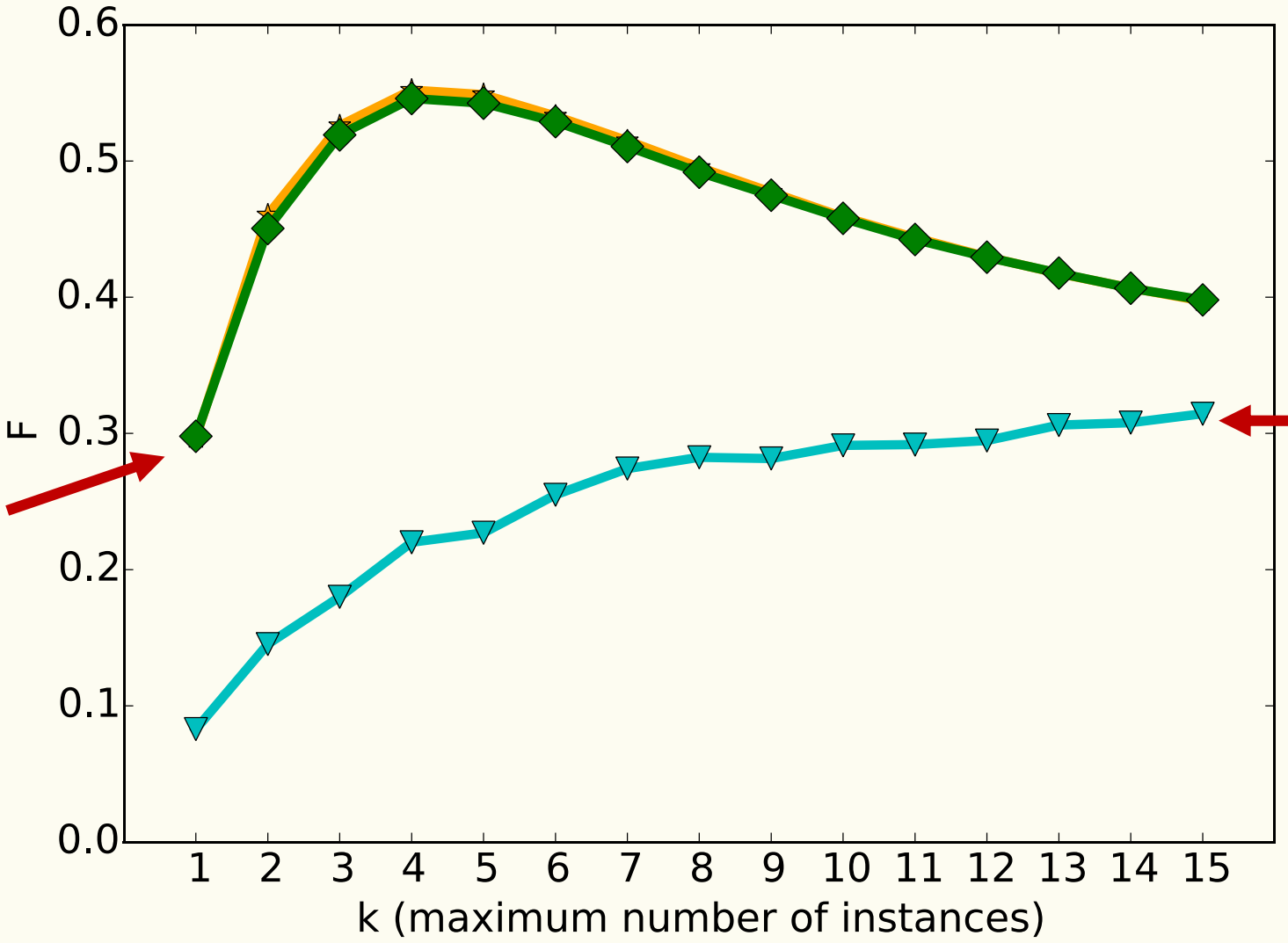
- Function Words
  - Random preposition/conjunction followed by optional '*the*'

[Woman]^{2} with [dress]^{0} and [boots]^{5} on the [car]^{3} .

Generating Descriptions: Results (F-score)



Generating Descriptions: Results (F-score)

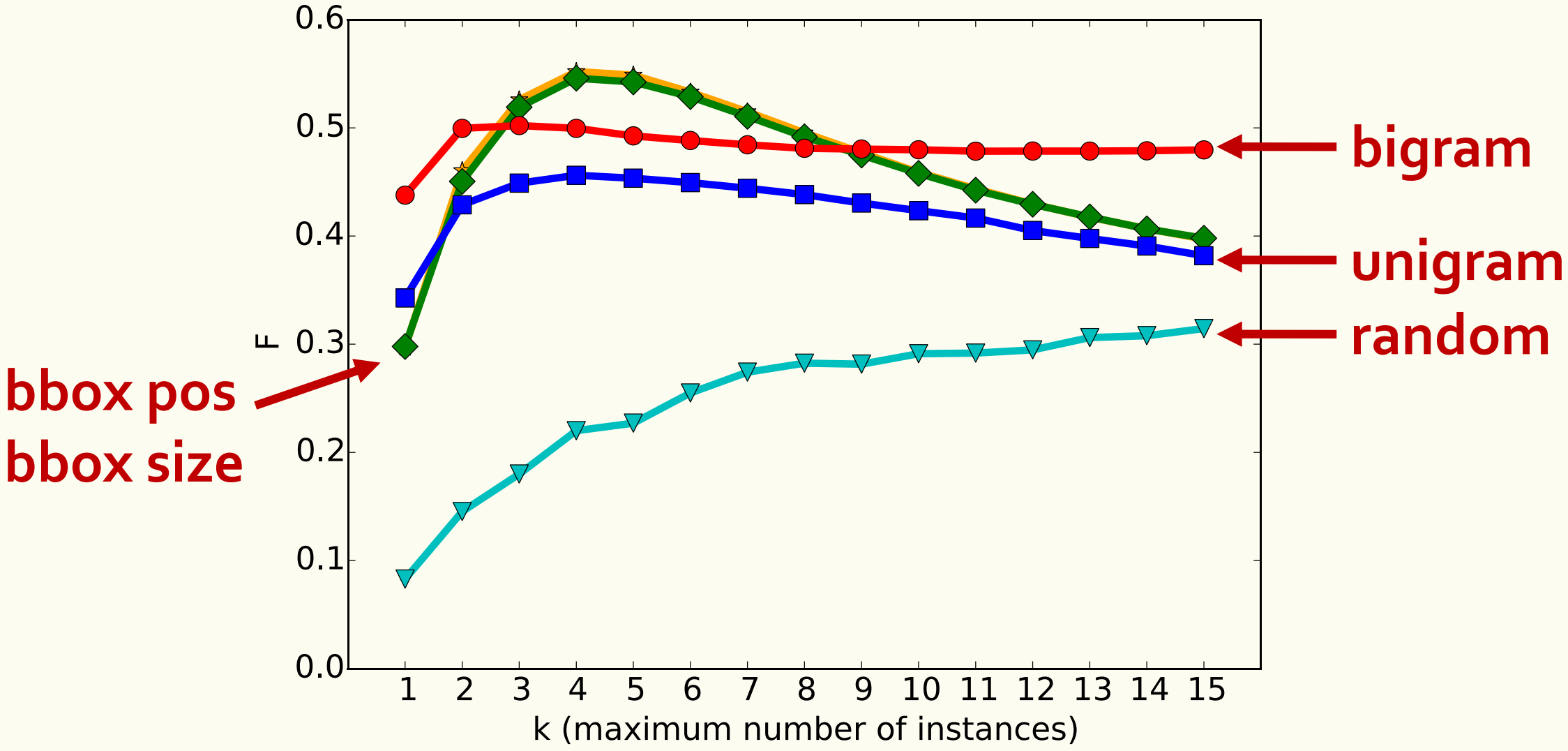


bbox pos

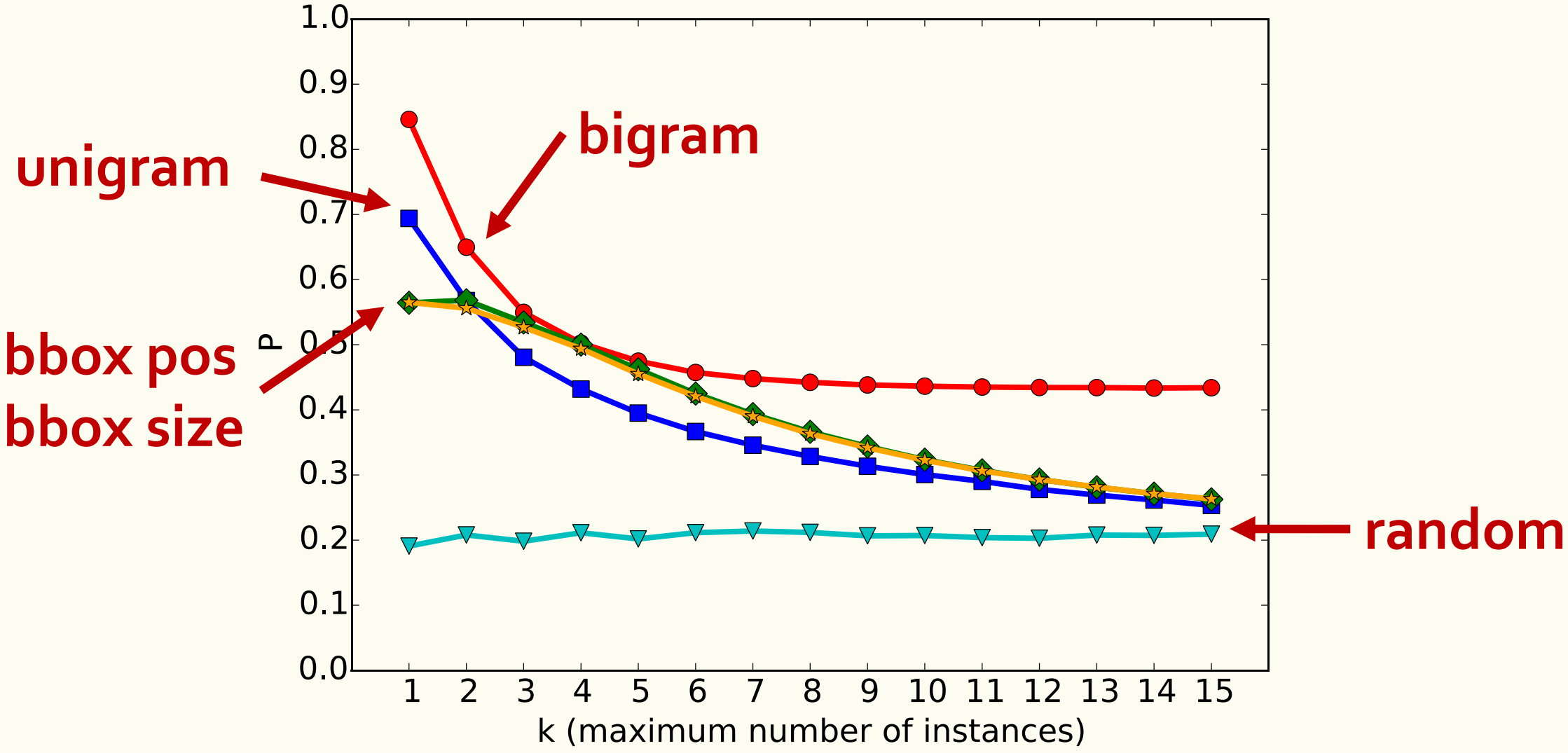
bbox size

random

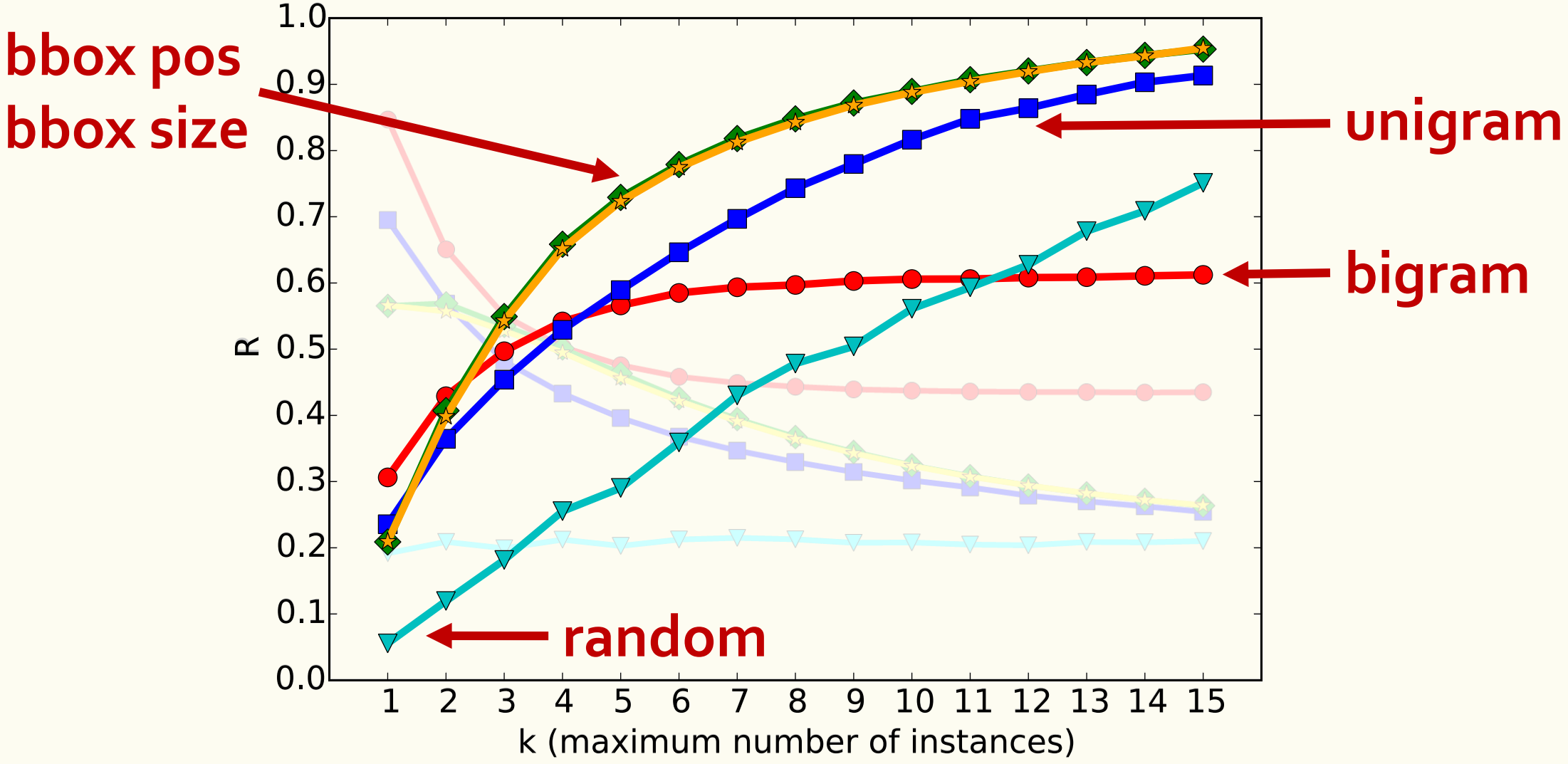
Generating Descriptions: Results (F-score)



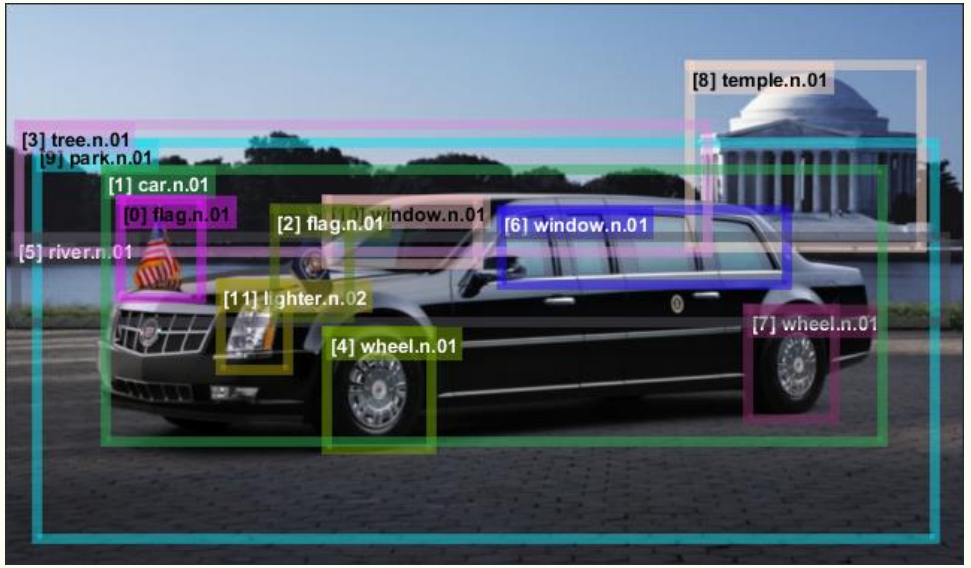
Generating Descriptions: Results (Precision)



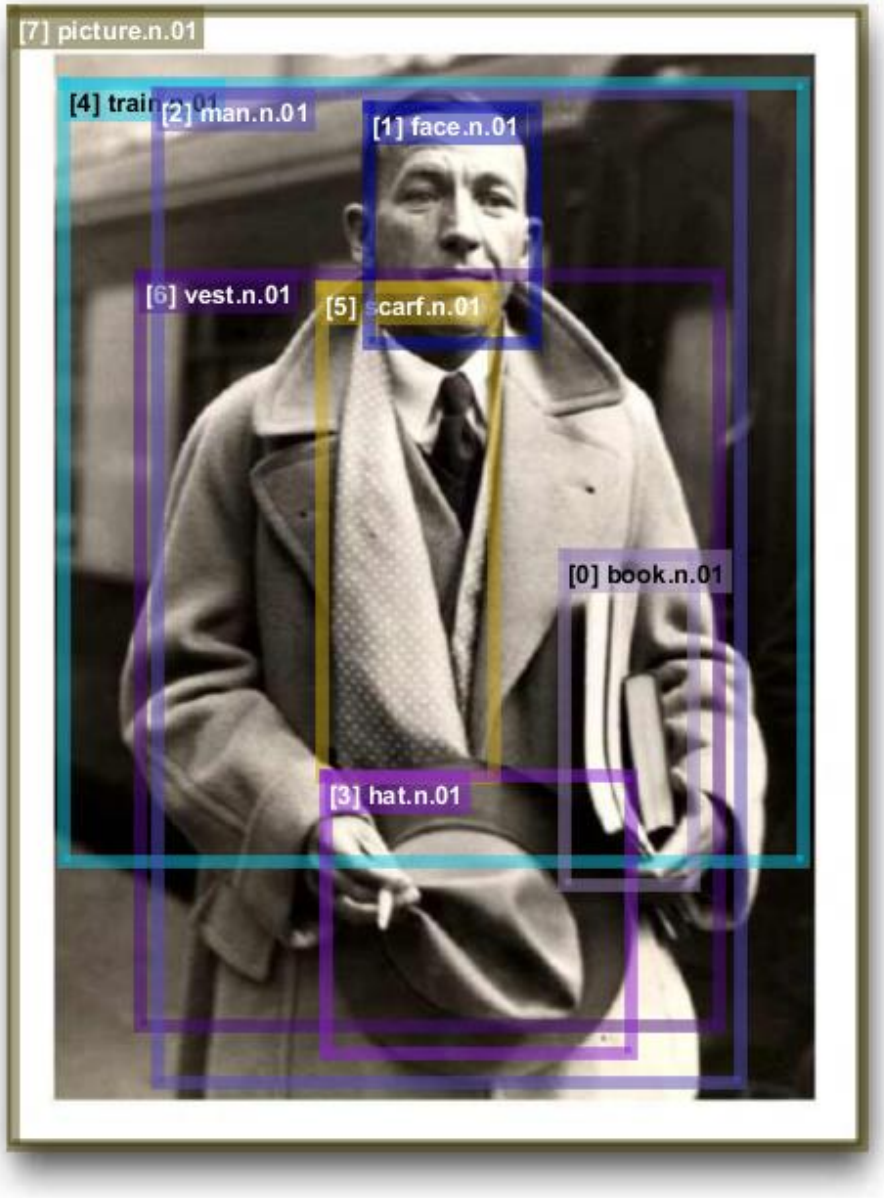
Generating Descriptions: Results (Recall)



- random [F: 0.43]
  - [Wall]^{4} inside [door]^{3} around the [bicycle]^{0} .
- bbox position [F: 0.79]
  - [Bicycle]^{0} in [floor]^{1} below [wall]^{2} .
- bbox size [F: 0.79]
  - [Bicycle]^{0} on [floor]^{1} with [wall]^{2} .
- unigram [F: 0.34]
  - [Table]^{7} in the [wall]^{4} around [wall]^{2} .
- bigram [F: 0.03]
  - [Table]^{7} near [door]^{3} .



- random [F: 0.05]
  - [Park][9] behind [wheel][7] underneath the [window][6] .
- bbox position [F: 0.59]
  - [Park][9] on the [car][1] below [river][5] .
- bbox size [F: 0.44]
  - [Park][9] behind the [car][1] against the [tree][3] .
- unigram [F: 0.42]
  - [Tree][3] beneath [car][1] by [window][6] .
- bigram [F: 0.71]
  - [Car][1] inside [flag][0] underneath the [flag][2] .



- random [F: 0.39]
  - [Vest]^{6} at [hat]^{3} behind the [picture]^{7} .
- bbox position [F: 0.49]
  - [Picture]^{7} on [man]^{2} beside the [train]^{4} .
- bbox size [F: 0.49]
  - [Picture]^{7} among [man]^{2} on the [train]^{4} .
- unigram [F: 0.77]
  - [Man]^{2} below the [hat]^{3} at [book]^{0} .
- bigram [F: 0.77]
  - [Man]^{2} around the [hat]^{3} along the [book]^{0} .

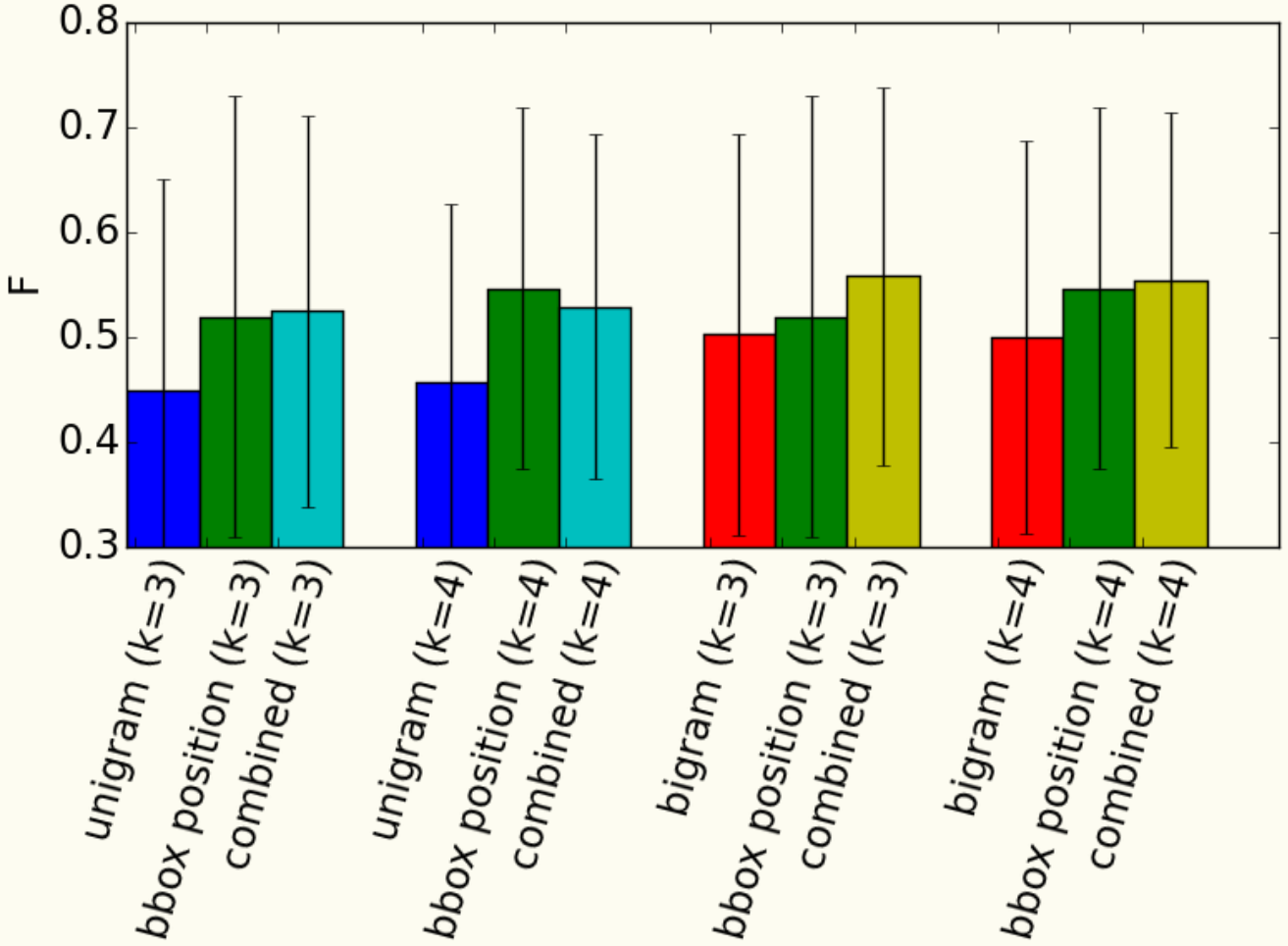
Generating Descriptions: Combined

- Combining Text + Visual Priors
  - Re-rank by average rank from two systems

	Bigram	Bbox Position	Average Rank	Combined
[0] dress	2	3	(2+3)/2 = 2.5	2
[1] wheel	- (6)	4	(6+4)/2 = 5	5
[2] woman	1	1	(1+1)/2 = 1	1
[3] car	4	2	(4+2)/2 = 3	3
[4] hair	- (6)	6	(6+6)/2 = 6	7
[5] boot	3	7	(3+7)/2 = 5	4
[6] sign	- (6)	5	(6+5)/2 = 5.5	6

[Woman]^{2} in [dress]^{0} by [car]^{3} . (for k=3)

Generating Descriptions: Combined



Discussion

- We presented the sentence generation task of ImageCLEF
  - Proposed content selection evaluation metric and baselines
- Challenges
  - Bounding box annotations may not be informative enough
  - Suitability of fine-grained metrics
- Future work
  - Fine-grained metrics
    - Content ordering, referring expressions, verbs/predicates/prepositions
  - Generation of image descriptions
    - Stronger cues (co-occurrences, spatial relations)

Generating Image Descriptions with Gold Standard Visual Inputs: Motivation, Evaluation and Baselines

Josiah Wang & Robert Gaizauskas

The University Of Sheffield.

Work presented as part of the Visual Sense (ViSen) consortium, funded by the chist-era D2K programme