Training Video Content in Natural Language Descriptions

Marcus Rohrbach, Wei Qiu, Ivan Titov,
Stefan Thater, Manfred Pinkal, Bernt Schiele

UNIVERSITY OF TORONTO

Outline

- Brief overview
- The 3 main questions of video description generation
- Data
- Different video description generation methods
- Brief intro to machine translation
- Technical approach
- Calculating BLEU
- Baselines, evaluations, and results
- Discussion

Overview

- Goal: finding natural language descriptions for video content
- Uses: Improvement of robotic interactions, generate summaries and descriptions for videos and movies, etc.

- Main contributions:
  1) Video description **phrased as a translation problem** from video content to natural language description, using the SR of the video content as an intermediate step.

  2) Approach evaluated on TACoS video description dataset.

  3) Annotations as well as intermediate outputs and final descriptions are **released on their website** - these allow for comparisons to their work or building on their SR

1) How to best approach the conversion from visual information to linguistic expressions?

- Use a two-step approach:

  -> Learn an intermediate Semantic Representation (SR) using a probabilistic model

  -> Given the SR, NLG problem phrased as *translation problem*, where the source is the SR and the target is a natural language description

2) Which part of the visual information is verbalized by humans and what is verbilized even though it is not directly present in the visual information?

- The most relevant information to verbalization and how to verbalize can be learnt from a parallel training corpus using SMT methods:

  a) Learn the correct ordering of words and phrases, referred to as surface realization in NLG

  b) Learn which SR should be realized in language

  c) Learn the proper correspondence between semantic concepts and verbalization, i.e. they do not have to define how semantic concepts are realized.

3) What is a good semantic representation (SR) of visual content and what is the limit of such a representation given perfect visual recognition?

- Compare three different visual representations
  - a raw video descriptor,
  - an attribute based representation,
  - the authors' CRF model.
- To understand the limits of their SR they also run the translation on ground truth annotations.

Data

- TACoS corpus of human-activity videos in a kitchen scenario
- People recorded preparing different kinds of ingredients
- Video lengths vary from 00:48 to 23:22
- TACoS parallel corpus contains a set of video snippets and sentences

Video sample from TACoS

Video's corresponding data

Video and data obtained from:

http://www.coli.uni-saarland.de/projects/smile/page.php?id=tacos

NLG from images and video

Four different ways of generating descriptions of visual content:

1) generating descriptions for (test) images or videos which already contain some associated text,

2) generating descriptions by using manually defined rules or templates,

3) retrieving existing descriptions from similar visual content,

4) learning a language model from a training corpus to generate descriptions.

Machine Translation

- For SMT you need:
  1) A language model:
     - P(Target text)
     - Used to generate fluent and grammatical output
       - Usually calculated using trigram statistics with back-off

  2) A translation model:
     - P(Target text | Source text)
       - Estimated based on sentence-aligned corpora of source and target languages

  3) A decoder:
     - Finds a sentence that maximizes the translation and language model probabilities
     - $T^* = \text{argmax}_T P(\text{Target text} \mid \text{Source text})\, P(\text{Target text})$.
- Moses (an open source toolkit) optimizes this pipeline on a training set.

Technical Approach: Overview

- x_i *:* Video snippets represented by the video descriptor
- z_i: a sentence
- (x_i, z_i): alignment
- (x_k, z_k) with x_k = x_i: if there is an extra descriptor for the same video snippet we treat it as an independent alignment
- y$_i$: intermediate level semantic representation (SR)
- y^*: SR for a new video (descriptor) x^*, predicted at test time.
- z^*: sentence generated from y^*.

Technical Approach: Overview (cont'd)

- Semantic Representation:
  - Based on the annotations provided with TACoS
  - Distinguishes *activities*, *tools*, *ingredients/objects*, *(source) location/container*, and (*target) location/container* in the form <ACTIVITY, TOOL, OBJECT, SOURCE, TARGET>.
  - NULL used for missing tool, object, or location
- SR annotations in TACoS can have a finer granularity than the sentences, i.e. (y_i^1, ...,y_i^{li}, ..., y_i^{Li}, z_i) where L_i is the number of SR annotations for sentence z_i
- For learning the SR extract the corresponding video snippet, i.e., (x_i^{li}, y_i^{li})
- No annotations at test time means no alignment problem when predicting y*

Ways of dealing with different levels of granularity

- For all SR annotations aligned to a sentence a separate training example is created, i.e. (y_i^1, z_i), ..., (y_i^{Li}, z_i).
- Only use the last SR (usually the most important one in TACoS) is used, i.e. (y_i^{Li}, z_i).
- Estimate the highest word overlap between the sentence and the string of the SR: $|y_i \cap Lemma(z_i)| / |y_i|$

  Lemma refers to lemmatizing, i.e., reducing to base forms

  e.g., *took* to *take*, *knives* to *knife, passed* to *pass*
- Predict one SR for each sentence, i.e. y_i* for z_i.

Technical Approach: Predicting a SR from visual content

1) Extract the visual content – different visual information usually highly correlated with each other
   E.g., activity *slice* more correlated with object *carrot* and tool *knife* than with *milk* and *spoon*

2) Model relationships with a Conditional Random Field (CRF). Visual entities modeled as nodes n_j observing the video descriptors *x* as unaries.

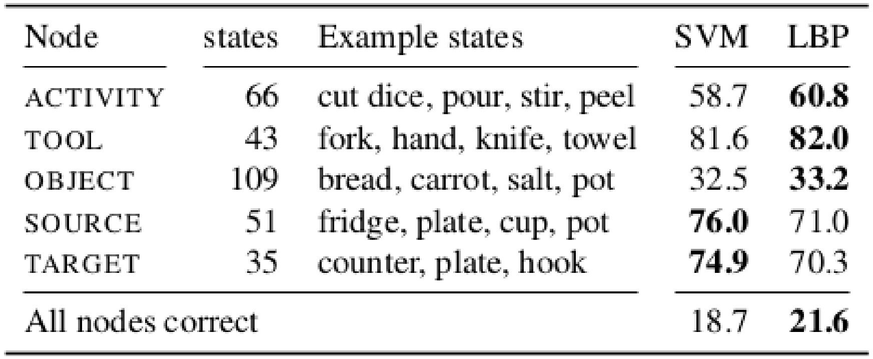
3) Graph is fully connected with learnt linear pairwise (p) and unary (u) weights using this standard energy formulation:

$$E(n_1, ..., n_N; x_i) = \sum_{j=1}^{N} E^u(n_j; x_i) + \sum_{j \sim k} E^p(n_j, n_k)$$

Technical Approach: Predicting a SR from visual content (cotn'd)

$$E(n_1, \ldots, n_N; x_i) = \sum_{j=1}^{N} E^u(n_j; x_i) + \sum_{j \sim k} E^p(n_j, n_k)$$

- $E^u(n_j;x_i) = \langle w_j^u, x_i \rangle$
- w^u_j: vector of the size of the video representation x_i
- $E^p(n_j,n_k) = w^p_{j,k}$
- Model learnt using training videos x_i^{li} and SR labels $y_i^{li} = \langle n_1, n_2, \ldots, n_N \rangle$ using loopy belief propagation (LBP)



Node	states	Example states	SVM	LBP
ACTIVITY	66	cut dice, pour, stir, peel	58.7	**60.8**
TOOL	43	fork, hand, knife, towel	81.6	**82.0**
OBJECT	109	bread, carrot, salt, pot	32.5	**33.2**
SOURCE	51	fridge, plate, cup, pot	**76.0**	71.0
TARGET	35	counter, plate, hook	**74.9**	70.3
All nodes correct			18.7	**21.6**

Table 1: CRF nodes of our SR. SVM vs. LBP inference: Node accuracy in % over all test sentences.

Technical Approach: Translating from a SR to a description

Converting SR to descriptions (SR -> D) is like translating from a source to a target language (L_S -> L_T)

Find the verbalization of a label n_i. e.g., HOB -> *stove*	Translate a word from L_S to L_T
Determine the ordering of the concepts of the SR in D	Find the alignment between two languages
Not necessarily all semantic concepts are verbalized in D. e.g., KNIFE not verbalized in *He cuts a carrot*	Certain words in L_S not represented in L_T or multiple words are combined to one. e.g., articles
Not necessarily all verbalized concepts are semantically represented. e.g, CUT, CARROT -> *He cuts the carrots*	Certain words in L_T not represented in L_S or one word becomes multiple
A language model of D is used to achieve a grammatically correct and fluent target sentence.	A language model of L_T is used to achieve a grammatically correct and fluent target sentence.

Technical Approach: Translating from a SR to a description (cont'd)

- SMT input: “*activity tool object source target*” where NULL states are converted to empty strings
- Giza++ learns an HMM concepts-word alignment model.
- This is the basis of the phrase-based translation model learned by Moses. Additionally a reordering model is learned based on the training data alignment statistics.
- IRSTLM estimates the fluency of the generated descriptions, based on n-gram statistics of TACoS.

Technical Approach: Translating from a SR to a description (cont'd)

- Optimize a linear model between the probabilities from the language model, phrase tables, and reordering model, as well as word, phrase, and rule counts.
- 10% of the training data is used as a validation set. In the optimization, BLEU @4 score used to compute the difference between predicted and provided reference descriptions.
- Testing: apply translation model to the SR *y** predicted by the CRF for a given input video *x**. This decoding results in the description *z**.

BLEU (BiLingual Evaluation Understudy) Score

- BLEU is a geometric mean over n-gram precisions
- Uses reference translation(s) and looks for local matches.
- Candidate sentences: machine-generated translation
- BLEU = $BP_C \times (p_1\ p_2\ p_3\ ...\ p_n)^{1/n}$
- p_n : the n-gram precision (e.g., BLEU @4 has n-gram precision of 4)
- BP: Brevity penalty; penalizes candidate sentence for having fewer words than the reference sentence(s)

Information from Frank Rudzicz's slides for the NLC course.

- Main flaw: A single sentence can be translated in many ways, with no overlap.
- However, in this experiment, the vocabulary is so constrained that this is O.K.

Baselines

- **Sentence retrieval:** Alternative to generating novel descriptions is to retrieve the first most likely sentence from a training corpus.
- **NLG with N-grams**: Keep the same SR but replace the SMT pipeline by learning a n-gram language model on the training set of descriptions.

  Basically predicts function words between SR-labels. For improved performance:

  1) Content words order identical in the target sentence;

  2) Tool and location frequently not verbalized => sensible string where only found when reduced to ACTIVITY and OBJECT;

  3) Only use the verb in the activity, e.g. CUT DICE -> *cut*, and the root word for noun phrases, e.g. PLASTIC BAG -> *bag*

Evaluation: Translating video to text

- 18,227 video/sentence pairs on 7,206 unique time intervals.
- 5609 intermediate level annotations, which form the SR (i.e., <ACTIVITY, TOOL, OBJECT, SOURCE, TARGET>).
- Dense trajectory features extract trajectory information, HOG, HOF, and MBH to form a descriptor of the video.
  => state-of-the-art performance on many activity recognition datasets, including TACoS.

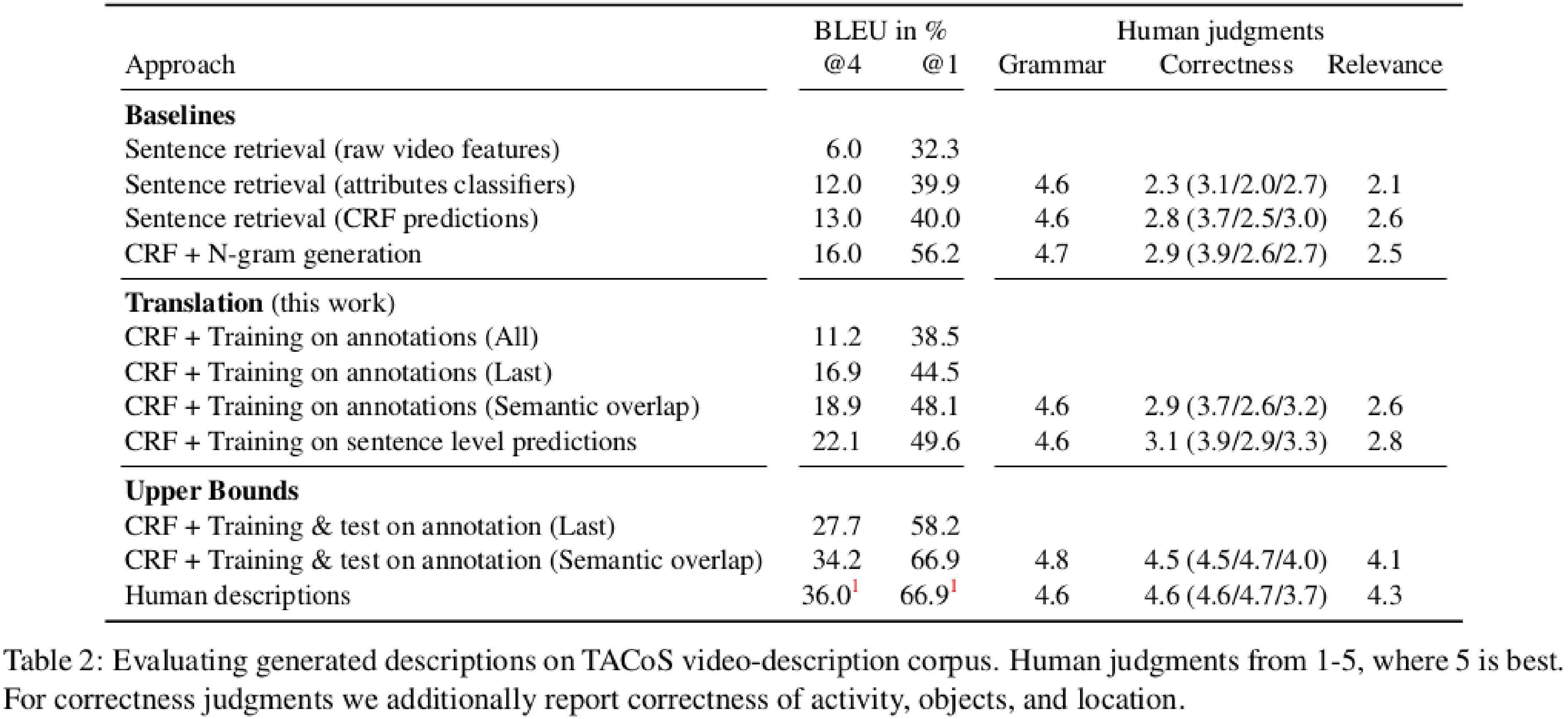
Evaluation: Translating video to text (cont'd)

- Tested on a subet of 490 video snippet / sentence pairs.
- CRF and Moses trained on the remaining TACoS corpus, using 10% as a validation set for parameter estimation.
- The attribute classifiers trained on the remaining videos of the MPII Cooking Composite Activity – a superset of TACoS.
- All text data preprocessed by substituting gender specific identifiers with “the person”

Evaluation: Translating video to text (cont'd)

- BLEU @4 (N=4) has shown to provide the best correlation with human judgements
- BLEU @1 provided for comparison with previous works' results
- For manual evaluation 10 human subjects rate:
  - grammatical correctness (independent of video content),
  - correctness (independent of grammatical correctness),
  - relevance (independent of grammatical correctness).
- Correctness: is the sentences correct with respect to the video?
- Relevance: does the sentence describe the most important activity and objects?
- Correctness of the activity, objects (tools and ingredients) separated from locations described.
- Scale from 1 to 5, with 5 = perfect, 1 = totaly bad.
- Continuous scores can be assigned (e.g., 3.5), if needed.
- Different sentences of the systems presented in a random order for each video.

Results: Translating video to text



Approach	BLEU in %		Human judgments		
	@4	@1	Grammar	Correctness	Relevance
Baselines					
Sentence retrieval (raw video features)	6.0	32.3			
Sentence retrieval (attributes classifiers)	12.0	39.9	4.6	2.3 (3.1/2.0/2.7)	2.1
Sentence retrieval (CRF predictions)	13.0	40.0	4.6	2.8 (3.7/2.5/3.0)	2.6
CRF + N-gram generation	16.0	56.2	4.7	2.9 (3.9/2.6/2.7)	2.5
Translation (this work)					
CRF + Training on annotations (All)	11.2	38.5			
CRF + Training on annotations (Last)	16.9	44.5			
CRF + Training on annotations (Semantic overlap)	18.9	48.1	4.6	2.9 (3.7/2.6/3.2)	2.6
CRF + Training on sentence level predictions	22.1	49.6	4.6	3.1 (3.9/2.9/3.3)	2.8
Upper Bounds					
CRF + Training & test on annotation (Last)	27.7	58.2			
CRF + Training & test on annotation (Semantic overlap)	34.2	66.9	4.8	4.5 (4.5/4.7/4.0)	4.1
Human descriptions	36.0[1]	66.9[1]	4.6	4.6 (4.6/4.7/3.7)	4.3

Table 2: Evaluating generated descriptions on TACoS video-description corpus. Human judgments from 1-5, where 5 is best. For correctness judgments we additionally report correctness of activity, objects, and location.

[1] Computed only on a 272 sentence subset where the corpus contains more than a single reference sentence for the same video. This reduces the number of references by one which leads to a lower BLEU score.

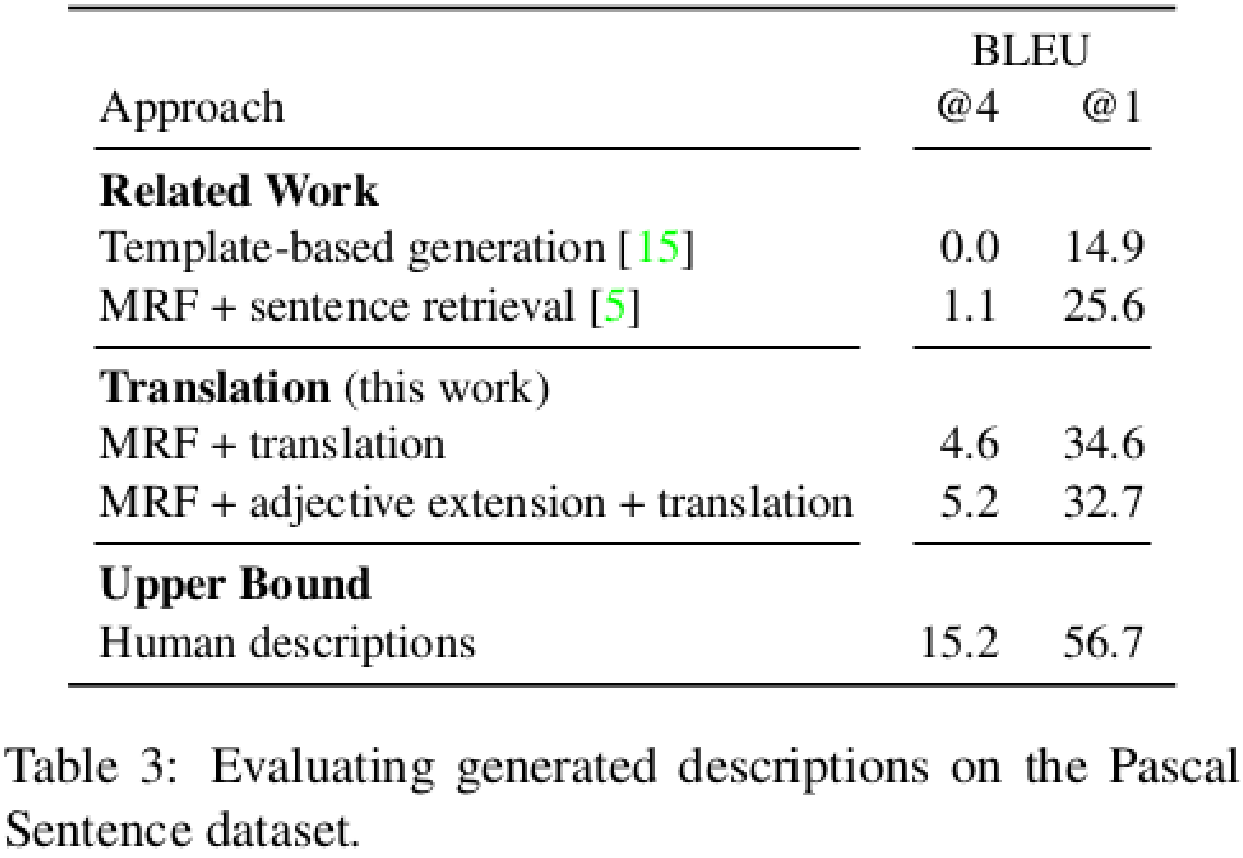
Results: Translating video to text (cont'd)

- Proposed approach using training on sentence level predictions outperforms all baselines
- Using the SR based on annotations very close to human performance (4.1 vs. 4.3, on a scale from 1 to 5, where 5 is the best).
- The grammatical correctness of the produced sentences disregarding the visual input: training and testing on annotations (score 4.8) outperforms the score for human descriptions (4.6), “indicating that our system learned a better language model than most human descriptions have.”
- Translation system achieves the same score as human descriptions.
- N-gram generation receives a slightly better score of 4.7 due to shorter sentences produced => fewer grammatical errors.

Evaluation: Translating images to text

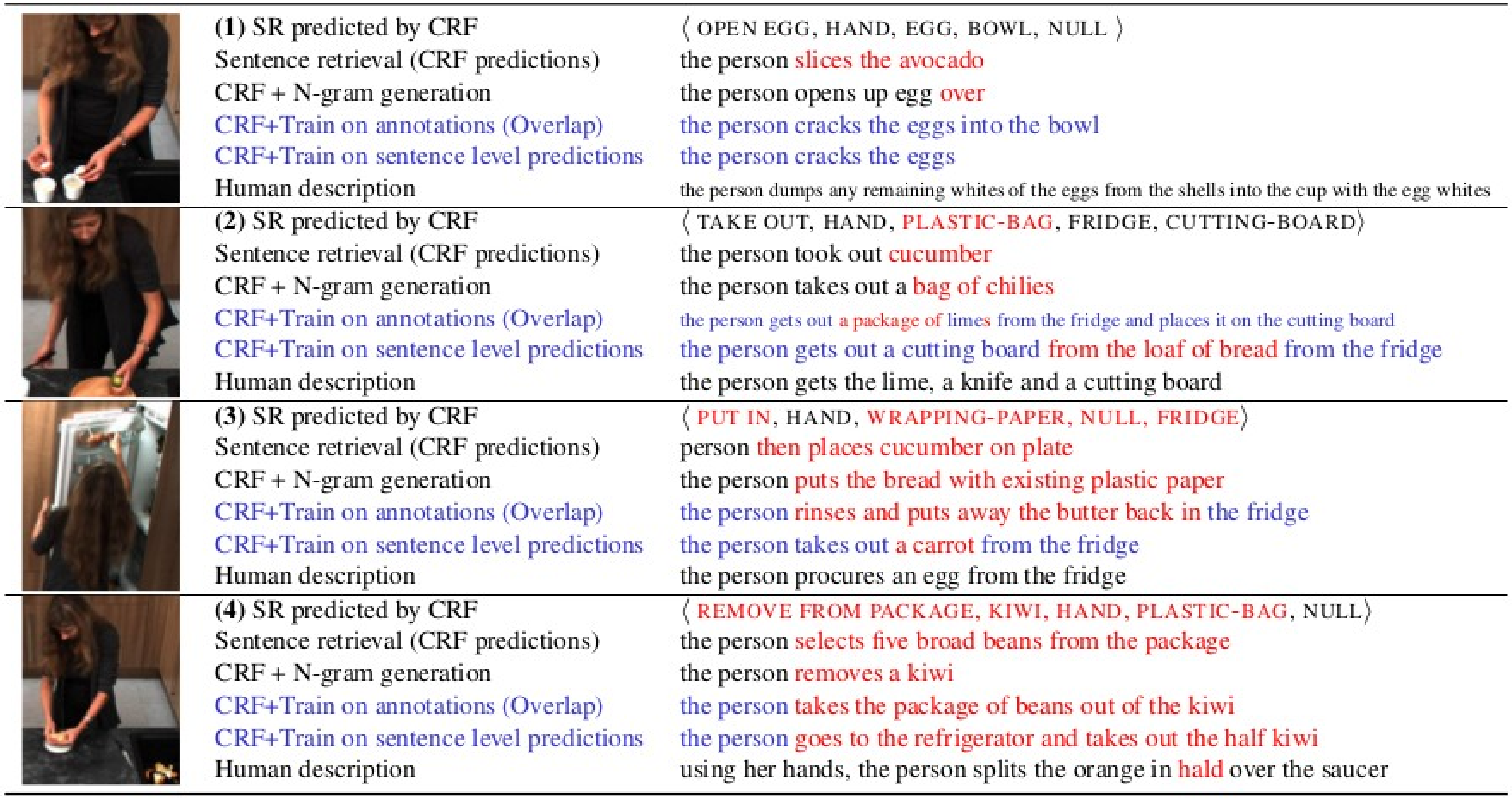
- Approach can applied to image decription.
- Use Pascal sentence dataset for evaluation (1,000 images, each paired with 5 different descriptions of one sentence), using the predictions provided by Farhadi et al. for the SR
- The SR consists of object-activity-scene triples
- Translation approach learnt on the training set of triples and image descriptions.
- Evaluated on a subset of 323 images with predicted descriptions from Farhadi et al. and Kulkarni et al.
- Use the first predicted triple (with highest score) from Farhadi et al.

Evaluation: Translating images to text (cont'd)



Approach	BLEU @4	BLEU @1
Related Work		
Template-based generation [15]	0.0	14.9
MRF + sentence retrieval [5]	1.1	25.6
Translation (this work)		
MRF + translation	4.6	34.6
MRF + adjective extension + translation	5.2	32.7
Upper Bound		
Human descriptions	15.2	56.7

Table 3: Evaluating generated descriptions on the Pascal Sentence dataset.



	(1) SR predicted by CRF	⟨ OPEN EGG, HAND, EGG, BOWL, NULL ⟩
	Sentence retrieval (CRF predictions)	the person slices the avocado
	CRF + N-gram generation	the person opens up egg over
	CRF+Train on annotations (Overlap)	the person cracks the eggs into the bowl
	CRF+Train on sentence level predictions	the person cracks the eggs
	Human description	the person dumps any remaining whites of the eggs from the shells into the cup with the egg whites
	(2) SR predicted by CRF	⟨ TAKE OUT, HAND, PLASTIC-BAG, FRIDGE, CUTTING-BOARD⟩
	Sentence retrieval (CRF predictions)	the person took out cucumber
	CRF + N-gram generation	the person takes out a bag of chilies
	CRF+Train on annotations (Overlap)	the person gets out a package of limes from the fridge and places it on the cutting board
	CRF+Train on sentence level predictions	the person gets out a cutting board from the loaf of bread from the fridge
	Human description	the person gets the lime, a knife and a cutting board
	(3) SR predicted by CRF	⟨ PUT IN, HAND, WRAPPING-PAPER, NULL, FRIDGE⟩
	Sentence retrieval (CRF predictions)	person then places cucumber on plate
	CRF + N-gram generation	the person puts the bread with existing plastic paper
	CRF+Train on annotations (Overlap)	the person rinses and puts away the butter back in the fridge
	CRF+Train on sentence level predictions	the person takes out a carrot from the fridge
	Human description	the person procures an egg from the fridge
	(4) SR predicted by CRF	⟨ REMOVE FROM PACKAGE, KIWI, HAND, PLASTIC-BAG, NULL⟩
	Sentence retrieval (CRF predictions)	the person selects five broad beans from the package
	CRF + N-gram generation	the person removes a kiwi
	CRF+Train on annotations (Overlap)	the person takes the package of beans out of the kiwi
	CRF+Train on sentence level predictions	the person goes to the refrigerator and takes out the half kiwi
	Human description	using her hands, the person splits the orange in hald over the saucer

Table 4: Example output of our system (blue) compared to baseline approaches and human descriptions, errors in red. (1, 2) our system provides the best output; (2, 3) our system partially recovers from a wrong SR; (4) failure case.

Discussion

- As the authors say, their work can be improved by:
  1) Modeling temporal dependencies in both the SR and the language generation
  2) Modeling the uncertainty of the visual input explicitely in the generation process
- SMT generation technique much more practical than a rule-based approach
- To improve, could make use of hypernyms and of classifying words as concrete (e.g., table) vs. abstract (e.g., freedom)