CBMM Memo No. 051 **September, 2015**

Do You See What I Mean? Visual Resolution of Linguistic Ambiguities

Published in the Proceedings of EMNLP 2015

Yevgeni Berzak, Andrei Barbu, Daniel Harari, Boris Katz and Shimon Ullman

Understanding language goes hand in hand with the ability to integrate complex contextual information obtained via perception. In this work, we present a novel task for grounded language understanding: disambiguating a sentence given a visual scene which depicts one of the possible interpretations of that sentence. To this end, we introduce a new multimodal corpus containing ambiguous sentences, representing a wide range of syntactic, semantic and discourse ambiguities, coupled with videos that visualize the different interpretations for each sentence. We address this task by extending a vision model which determines if a sentence is depicted by a video. We demonstrate how such a model can be adjusted to recognize different interpretations of the same underlying sentence, allowing to disambiguate sentences in a unified fashion across the different ambiguity types.

This work was supported by the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF-1231216.

Do You See What I Mean?
Visual Resolution of Linguistic Ambiguities

Yevgeni Berzak
CSAIL MIT
berzak@mit.edu

Andrei Barbu
CSAIL MIT
andrei@0xab.com

Daniel Harari
CSAIL MIT
hararid@mit.edu

Boris Katz
CSAIL MIT
boris@mit.edu

Shimon Ullman
Weizmann Institute of Science
shimon.ullman@weizmann.ac.il

Abstract

Understanding language goes hand in hand with the ability to integrate complex contextual information obtained via perception. In this work, we present a novel task for grounded language understanding: disambiguating a sentence given a visual scene which depicts one of the possible interpretations of that sentence. To this end, we introduce a new multimodal corpus containing ambiguous sentences, representing a wide range of syntactic, semantic and discourse ambiguities, coupled with videos that visualize the different interpretations for each sentence. We address this task by extending a vision model which determines if a sentence is depicted by a video. We demonstrate how such a model can be adjusted to recognize different interpretations of the same underlying sentence, allowing to disambiguate sentences in a unified fashion across the different ambiguity types.

1 Introduction

Ambiguity is one of the defining characteristics of human languages, and language understanding crucially relies on the ability to obtain unambiguous representations of linguistic content. While some ambiguities can be resolved using intra-linguistic contextual cues, the disambiguation of many linguistic constructions requires integration of world knowledge and perceptual information obtained from other modalities.

In this work, we focus on the problem of grounding language in the visual modality, and introduce a novel task for language understanding which requires resolving linguistic ambiguities by utilizing the visual context in which the linguistic content is expressed. This type of inference is frequently called for in human communication that occurs in a visual environment, and is crucial for language acquisition, when much of the linguistic content refers to the visual surroundings of the child (Snow, 1972).

Our task is also fundamental to the problem of grounding vision in language, by focusing on phenomena of linguistic ambiguity, which are prevalent in language, but typically overlooked when using language as a medium for expressing understanding of visual content. Due to such ambiguities, a superficially appropriate description of a visual scene may in fact not be sufficient for demonstrating a correct understanding of the relevant visual content. Our task addresses this issue by introducing a deep validation protocol for visual understanding, requiring not only providing a surface description of a visual activity but also demonstrating structural understanding at the levels of syntax, semantics and discourse.

To enable the systematic study of visually grounded processing of ambiguous language, we create a new corpus, *LAVA* (Language and Vision Ambiguities). This corpus contains sentences with linguistic ambiguities that can only be resolved using external information. The sentences are paired with short videos that visualize different interpretations of each sentence. Our sentences encompass a wide range of syntactic, semantic and dis-

Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 1477–1487,
Lisbon, Portugal, 17-21 September 2015. ©2015 Association for Computational Linguistics.

course ambiguities, including ambiguous prepositional and verb phrase attachments, conjunctions, logical forms, anaphora and ellipsis. Overall, the corpus contains 237 sentences, with 2 to 3 interpretations per sentence, and an average of 3.37 videos that depict visual variations of each sentence interpretation, corresponding to a total of 1679 videos.

Using this corpus, we address the problem of selecting the interpretation of an ambiguous sentence that matches the content of a given video. Our approach for tackling this task extends the *sentence tracker* introduced in (Siddharth et al., 2014). The sentence tracker produces a score which determines if a sentence is depicted by a video. This earlier work had no concept of ambiguities; it assumed that every sentence had a single interpretation. We extend this approach to represent multiple interpretations of a sentence, enabling us to pick the interpretation that is most compatible with the video.

To summarize, the contributions of this paper are threefold. First, we introduce a new task for visually grounded language understanding, in which an ambiguous sentence has to be disambiguated using a visual depiction of the sentence's content. Second, we release a multimodal corpus of sentences coupled with videos which covers a wide range of linguistic ambiguities, and enables a systematic study of linguistic ambiguities in visual contexts. Finally, we present a computational model which disambiguates the sentences in our corpus with an accuracy of 75.36%.

2 Related Work

Previous language and vision studies focused on the development of multimodal word and sentence representations (Bruni et al., 2012; Socher et al., 2013; Silberer and Lapata, 2014; Gong et al., 2014; Lazaridou et al., 2015), as well as methods for describing images and videos in natural language (Farhadi et al., 2010; Kulkarni et al., 2011; Mitchell et al., 2012; Socher et al., 2014; Thomason et al., 2014; Karpathy and Fei-Fei, 2014; Siddharth et al., 2014; Venugopalan et al., 2015; Vinyals et al., 2015). While these studies handle important challenges in multimodal processing of language and vision, they do not provide explicit modeling of linguistic ambiguities.

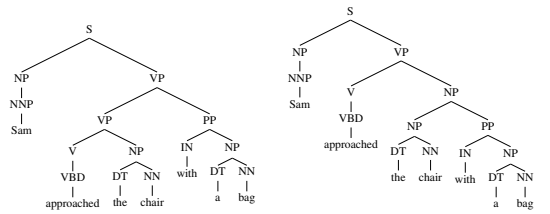
Previous work relating ambiguity in language to the visual modality addressed the problem of word sense disambiguation (Barnard et al., 2003). However, this work is limited to context independent interpretation of individual words, and does not consider structure-related ambiguities. Discourse ambiguities were previously studied in work on multimodal coreference resolution (Ramanathan et al., 2014; Kong et al., 2014). Our work expands this line of research, and addresses further discourse ambiguities in the interpretation of ellipsis. More importantly, to the best of our knowledge our study is the first to present a systematic treatment of syntactic and semantic sentence level ambiguities in the context of language and vision.

The interactions between linguistic and visual information in human sentence processing have been extensively studied in psycholinguistics and cognitive psychology (Tanenhaus et al., 1995). A considerable fraction of this work focused on the processing of ambiguous language (Spivey et al., 2002; Coco and Keller, 2015), providing evidence for the importance of visual information for linguistic ambiguity resolution by humans. Such information is also vital during language acquisition, when much of the linguistic content perceived by the child refers to their immediate visual environment (Snow, 1972). Over time, children develop mechanisms for grounded disambiguation of language, manifested among others by the usage of iconic gestures when communicating ambiguous linguistic content (Kidd and Holler, 2009). Our study leverages such insights to develop a complementary framework that enables addressing the challenge of visually grounded disambiguation of language in the realm of artificial intelligence.

3 Task

In this work we provide a concrete framework for the study of language understanding with visual context by introducing the task of grounded language disambiguation. This task requires to choose the correct linguistic representation of a sentence given a visual context depicted in a video. Specifically, provided with a sentence, n candidate interpretations of that sentence and a video that depicts the content of the sentence, one needs to choose the interpretation that corresponds to the content of the video.

To illustrate this task, consider the example in figure 1, where we are given the sentence "Sam approached the chair with a bag" along with two different linguistic interpretations. In the first in-

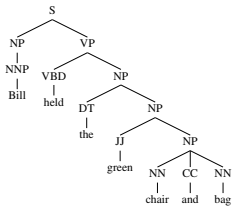


(a) First interpretation

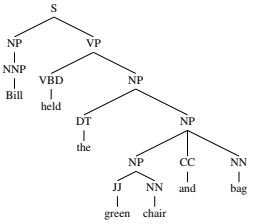
(b) Second interpretation

(c) Visual context

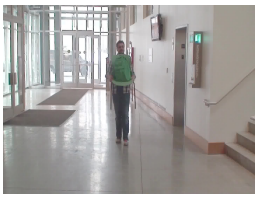
Figure 1: An example of the visually grounded language disambiguation task. Given the sentence "Sam approached the chair with a bag", two potential parses, (a) and (b), correspond to two different semantic interpretations. In the first interpretation Sam has the bag, while in the second reading the bag is on the chair. The task is to select the correct interpretation given the visual context (c).



(a) First interpretation



(b) Second interpretation



(c) First visual context

(d) Second visual context

Figure 2: Linguistic and visual interpretations of the sentence "Bill held the green chair and bag". In the first interpretation (a,c) both the chair and bag are green, while in the second interpretation (b,d) only the chair is green and the bag has a different color.

terpretation, which corresponds to parse 1(a), Sam has the bag. In the second interpretation associated with parse 1(b), the bag is on the chair rather than with Sam. Given the visual context from figure 1(c), the task is to choose which interpretation is most appropriate for the sentence.

4 Approach Overview

To address the grounded language disambiguation task, we use a compositional approach for determining if a specific interpretation of a sentence is depicted by a video. In this framework, described in detail in section 6, a sentence and an accompanying interpretation encoded in first order logic, give rise to a grounded model that matches a video against the provided sentence interpretation.

The model is comprised of Hidden Markov Models (HMMs) which encode the semantics of words, and trackers which locate objects in video frames. To represent an interpretation of a sentence, word models are combined with trackers through a cross-product which respects the semantic representation of the sentence to create a single model which recognizes that interpretation.

Given a sentence, we construct an HMM based representation for each interpretation of that sentence. We then detect candidate locations for objects in every frame of the video. Together the reforestation for the sentence and the candidate object locations are combined to form a model which can determine if a given interpretation is depicted by the video. We test each interpretation and report the interpretation with highest likelihood.

5 Corpus

To enable a systematic study of linguistic ambiguities that are grounded in vision, we compiled a corpus with ambiguous sentences describing visual actions. The sentences are formulated such that the correct linguistic interpretation of each sentence can only be determined using external, non-linguistic, information about the depicted activity. For example, in the sentence "Bill held the green chair and bag", the correct scope of "green" can only be determined by integrating additional information about the color of the bag. This information is provided in the accompanying videos, which visualize the possible interpretations of each sentence. Figure 2 presents the syntactic parses for this example along with frames from the respective videos. Although our videos contain visual uncertainty, they are not ambiguous with respect to the linguistic interpretation they are presenting, and hence a video always corresponds to a single candidate representation of a sentence.

The corpus covers a wide range of well

known syntactic, semantic and discourse ambiguity classes. While the ambiguities are associated with various types, different sentence interpretations always represent distinct sentence meanings, and are hence encoded semantically using first order logic. For syntactic and discourse ambiguities we also provide an additional, ambiguity type specific encoding as described below.

- **Syntax** Syntactic ambiguities include Prepositional Phrase (PP) attachments, Verb Phrase (VP) attachments, and ambiguities in the interpretation of conjunctions. In addition to logical forms, sentences with syntactic ambiguities are also accompanied with Context Free Grammar (CFG) parses of the candidate interpretations, generated from a deterministic CFG parser.

- **Semantics** The corpus addresses several classes of semantic quantification ambiguities, in which a syntactically unambiguous sentence may correspond to different logical forms. For each such sentence we provide the respective logical forms.

- **Discourse** The corpus contains two types of discourse ambiguities, Pronoun Anaphora and Ellipsis, offering examples comprising two sentences. In **anaphora** ambiguity cases, an ambiguous pronoun in the second sentence is given its candidate antecedents in the first sentence, as well as a corresponding logical form for the meaning of the second sentence. In **ellipsis** cases, a part of the second sentence, which can constitute either the subject and the verb, or the verb and the object, is omitted. We provide both interpretations of the omission in the form of a single unambiguous sentence, and its logical form, which combines the meanings of the first and the second sentences.

Table 2 lists examples of the different ambiguity classes, along with the candidate interpretations of each example.

The corpus is generated using Part of Speech (POS) tag sequence templates. For each template, the POS tags are replaced with lexical items from the corpus lexicon, described in table 3, using all the visually applicable assignments. This generation process yields an overall of 237 sentences,

	Ambiguity	Templates	#
Syntax	PP	NNP V DT [JJ] NN_1 IN DT [JJ] NN_2.	48
	VP	NNP_1 V [IN] NNP_2 V [JJ] NN.	60
	Conjunction	NNP_1 [and NNP_2] V DT JJ NN_1 and NN_2. NNP V DT NN_1 or DT NN_2 and DT NN_3.	40
Semantics	Logical Form	NNP_1 and NNP_2 V a NN. Someone V the NNS.	35
Discourse	Anaphora	NNP V DT NN_1 and DT NN_2. It is JJ.	36
	Ellipsis	NNP_1 V NNP_2. Also NNP_3.	18

Table 1: POS templates for generating the sentences in our corpus. The rightmost column represents the number of sentences in each category. The sentences are produced by replacing the POS tags with all the visually applicable assignments of lexical items from the corpus lexicon shown in table 3.

of which 213 sentences have 2 candidate interpretations, and 24 sentences have 3 interpretations. Table 1 presents the corpus templates for each ambiguity class, along with the number of sentences generated from each template.

The corpus videos are filmed in an indoor environment containing background objects and pedestrians. To account for the manner of performing actions, videos are shot twice with different actors. Whenever applicable, we also filmed the actions from two different directions (e.g. approach from the left, and approach from the right). Finally, all videos were shot with two cameras from two different view points. Taking these variations into account, the resulting video corpus contains 7.1 videos per sentence and 3.37 videos per sentence interpretation, corresponding to a total of 1679 videos. The average video length is 3.02 seconds (90.78 frames), with in an overall of 1.4 hours of footage (152434 frames).

A custom corpus is required for this task because no existing corpus, containing either videos or images, systematically covers multimodal ambiguities. Datasets such as *UCF Sports* (Rodriguez et al., 2008), *YouTube* (Liu et al., 2009), and *HMDB* (Kuehne et al., 2011) which come out of the activity recognition community are accompanied by action labels, not sentences, and do not control for the content of the videos aside from the principal action being performed. Datasets for image and video captioning, such as *MSCOCO* (Lin et al., 2014) and *TACOS* (Regneri et al., 2013),

	Ambiguity	Example	Linguistic interpretations	Visual setups
Syntax	PP	Claire left the green chair with a yellow bag.	Claire [left the green chair] [with a yellow bag].	The bag is with Claire.
			Claire left [the green chair with a yellow bag].	Bag is on the chair.
	VP	Claire looked at Bill picking up a chair.	Claire looked at [Bill [picking up a chair]].	Bill picks up the chair.
			Claire [looked at Bill] [picking up a chair].	Claire picks up the chair.
	Conjunction	Claire held a green bag and chair.	Claire held a [green [bag and chair]].	The chair is green.
			Claire held a [[green bag] and [chair]].	The chair is not green.
		Claire held the chair or the bag and the telescope.	Claire held [[the chair] or [the bag and the telescope]].	Claire holds the chair.
			Claire held [[the chair or the bag] and [the telescope]].	Claire holds the chair and the telescope.
Semantics	Logical Form	Claire and Bill moved a chair.	**chair**(x), **move**(Claire, x), **move**(Bill, x)	Claire and Bill move the same chair.
			chair(x), **chair**(y), **move**(Claire, x), **move**(Bill, y), $x \neq y$	Claire and Bill move different chairs.
		Someone moved the two chairs.	**chair**(x), **chair**(y), $x \neq y$, **person**(u), **move**(u, x), **move**(u, y)	One person moves both chairs.
			chair(x), **chair**(y), $x \neq y$, **person**(u), **person**(v), $u \neq v$, **move**(u, x), **move**(v, y)	Each chair moved by a different person.
Discourse	Anaphora	Claire held the bag and the chair. It is yellow.	It = bag	The bag is yellow.
			It = chair	The chair is yellow.
	Ellipsis	Claire looked at Bill. Also Sam.	Claire looked at Bill and Sam.	Claire looks at Bill and Sam.
			Claire and Sam looked at Bill.	Claire and Sam look at Bill.

Table 2: An overview of the different ambiguity types, along with examples of ambiguous sentences with their linguistic and visual interpretations. Note that similarly to semantic ambiguities, syntactic and discourse ambiguities are also provided with first order logic formulas for the resulting sentence interpretations. Table 4 shows additional examples for each ambiguity type, with frames from sample videos corresponding to the different interpretations of each sentence.

Syntactic Category	Visual Category	Words
Nouns	Objects, People	chair, bag, telescope, someone, proper names
Verbs	Actions	pick up, put down, hold, move (transitive), look at, approach, leave
Prepositions	Spacial Relations	with, left of, right of, on
Adjectives	Visual Properties	yellow, green

Table 3: The lexicon used to instantiate the templates in figure 1 in order to generate the corpus.

aim to control for more aspects of the videos than just the main action being performed but they do not provide the range of ambiguities discussed here. The closest dataset is that of Siddharth et al. (2014) as it controls for object appearance, color, action, and direction of motion, making it more likely to be suitable for evaluating disambiguation tasks. Unfortunately, that dataset was designed to avoid ambiguities, and therefore is not suitable for evaluating the work described here.

6 Model

To perform the disambiguation task, we extend the sentence recognition model of Siddharth et al. (2014) which represents sentences as compositions of words. Given a sentence, its first order logic interpretation and a video, our model produces a score which determines if the sentence is depicted by the video. It simultaneously tracks the participants in the events described by the sentence while recognizing the events themselves. This allows it to be flexible in the presence of noise by integrating top-down information from the sentence with bottom-up information from object and property detectors. Each word in the query sentence is represented by an HMM (Baum et al., 1970), which recognizes tracks (i.e. paths of detections in a video for a specific object) that satisfy the semantics of the given word. In essence, this model can be described as having two layers, one in which object tracking occurs and one in which words observe tracks and filter tracks that do not satisfy the word constraints.

Given a sentence interpretation, we construct a sentence-specific model which recognizes if a video depicts the sentence as follows. Each predicate in the first order logic formula has a corresponding HMM, which can recognize if that predicate is true of a video given its arguments. Each variable has a corresponding tracker which attempts to physically locate the bounding box corresponding to that variable in each frame of a

PP Attachment	Sam looked at Bill with a telescope.
VP Attachment	Bill approached the person holding a green chair.
Conjunction	Sam and Bill picked up the yellow bag and chair.
Logical Form	Someone put down the bags.
Anaphora	Sam picked up the bag and the chair. It is yellow.
Ellipsis	Sam left Bill. Also Clark.

Table 4: Examples of the six ambiguity classes described in table 2. The example sentences have at least two interpretations, which are depicted by different videos. Three frames from each such video are shown on the left and on the right below each sentence.

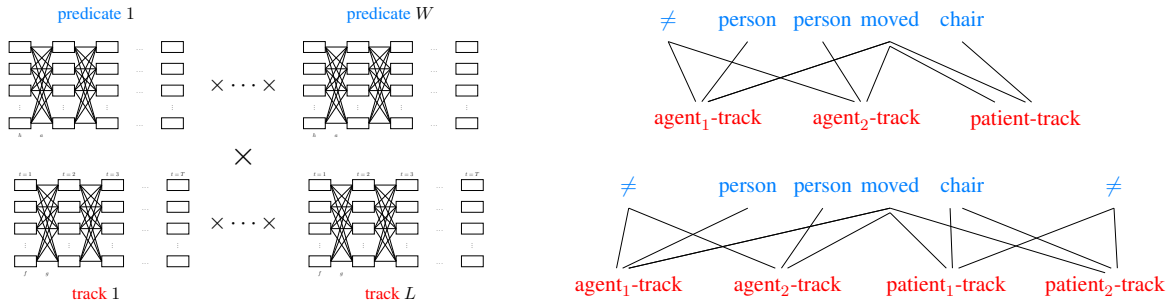


Figure 3: (left) Tracker lattices for every sentence participant are combined with predicate HMMs. The MAP estimate in the resulting cross-product lattice simultaneously finds the best tracks and the best state sequences for every predicate. (right) Two interpretations of the sentence "Claire and Bill moved a chair" having different first order logic formulas. The top interpretation corresponds to Bill and Claire moving the same chair, while the bottom one describes them moving different chairs. Predicates are highlighted in blue at the top and variables are highlighted in red at the bottom. Each predicate has a corresponding HMM which recognizes its presence in a video. Each variable has a corresponding tracker which locates it in a video. Lines connect predicates and the variables which fill their argument slots. Some predicates, such as *move* and $\neq$, take multiple arguments. Some predicates, such as *move*, are applied multiple times between different pairs of variables.

video. This creates a bipartite graph: HMMs that represent predicates are connected to trackers that represent variables. The trackers themselves are similar to the HMMs, in that they comprise a lattice of potential bounding boxes in every frame. To construct a joint model for a sentence interpretation, we take the cross product of HMMs and trackers, taking only those cross products dictated by the structure of the formula corresponding to the desired interpretation. Given a video, we employ an object detector to generate candidate detections in each frame, construct trackers which select one of these detections in each frame, and finally construct the overall model from HMMs and trackers.

Provided an interpretation and its corresponding formula composed of P predicates and V variables, along with a collection of object detections, $b^{\text{frame}}_{\text{detection index}}$, in each frame of a video of length T the model computes the score of the video-sentence pair by finding the optimal detection for each participant in every frame. This is in essence the Viterbi algorithm (Viterbi, 1971), the MAP algorithm for HMMs, applied to finding optimal object detections $j^{\text{frame}}_{\text{variable}}$ for each participant, and the optimal state $k^{\text{frame}}_{\text{predicate}}$ for each predicate HMM, in every frame. Each detection is scored by its confidence from the object detector, f and each object track is scored by a motion coherence metric g which determines if the motion of the track agrees with the underlying optical flow. Each predicate, p, is scored by the probability of observing a particular detection in a given state h_p, and by the probability of transitioning between states a_p. The structure of the formula and the fact that multiple predicates often refer to the same variables is recorded by θ, a mapping between predicates and their arguments. The model computes the MAP estimate as:

$$\max_{j_1^1,\ldots,j_1^T} \cdots \max_{j_V^1,\ldots,j_V^T} \max_{k_1^1,\ldots,k_1^T} \cdots \max_{k_P^1,\ldots,k_P^T} \sum_{v=1}^{V}\sum_{t=1}^{T} f(b^t_{j_v^t}) + \sum_{t=2}^{T} g(b^{t-1}_{j_v^{t-1}}, b^t_{j_v^t}) + \sum_{p=1}^{P}\sum_{t=1}^{T} h_p(k_p^t, b^t_{j^t_{\theta_p^1}}, b^t_{j^t_{\theta_p^2}}) + \sum_{t=2}^{T} a_p(k_p^{t-1}, k_p^t)$$

for sentences which have words that refer to at most two tracks (i.e. transitive verbs or binary predicates) but is trivially extended to arbitrary arities. Figure 3 provides a visual overview of the model as a cross-product of tracker models and word models.

Our model extends the approach of Siddharth et al. (2014) in several ways. First, we depart from the dependency based representation used in that work, and recast the model to encode first order logic formulas. Note that some complex first order logic formulas cannot be directly encoded in the model and require additional inference steps. This extension enables us to represent ambiguities in which a given sentence has multiple logical interpretations for the same syntactic parse.

Second, we introduce several model components which are not specific to disambiguation, but are required to encode linguistic constructions that are present in our corpus and could not be handled by the model of Siddharth et al. (2014). These new components are the predicate "not equal", disjunction, and conjunction. The key addition among these components is support for the new predicate "not equal", which enforces that two tracks, i.e. objects, are distinct from each other. For example, in the sentence "Claire and Bill moved a chair" one would want to ensure that the two movers are distinct entities. In earlier work, this was not required because the sentences tested in that work were designed to distinguish objects based on constraints rather than identity. In other words, there might have been two different people but they were distinguished in the sentence by their actions or appearance. To faithfully recognize that two actors are moving the chair in the earlier example, we must ensure that they are disjoint from each other. In order to do this we create a new HMM for this predicate, which assigns low probability to tracks that heavily overlap, forcing the model to fit two different actors in the previous example. By combining the new first order logic based semantic representation in lieu of a syntactic representation with a more expressive model, we can encode the sentence interpretations required to perform the disambiguation task.

Figure 3(left) shows an example of two different interpretations of the above discussed sentence "Claire and Bill moved a chair". Object trackers, which correspond to variables in the first order logic representation of the sentence interpretation, are shown in red. Predicates which constrain the possible bindings of the trackers, corresponding to predicates in the representation of the sentence, are shown in blue. Links represent the argument structure of the first order logic formula, and determine the cross products that are taken between the predicate HMMs and tracker lattices in order to form the joint model which recognizes the entire interpretation in a video.

The resulting model provides a single unified formalism for representing all the ambiguities in table 2. Moreover, this approach can be tuned to different levels of specificity. We can create models that are specific to one interpretation of a sentence or that are generic, and accept multiple interpretations by eliding constraints that are not common between the different interpretations. This allows the model, like humans, to defer deciding on a particular interpretation or to infer that multiple interpretation of the sentence are plausible.

7 Experimental Results

We tested the performance of the model described in the previous section on the *LAVA* dataset presented in section 5. Each video in the dataset was pre-processed with object detectors for humans, bags, chairs, and telescopes. We employed a mixture of CNN (Krizhevsky et al., 2012) and DPM (Felzenszwalb et al., 2010) detectors, trained on held out sections of our corpus. For each object class we generated proposals from both the CNN and the DPM detectors, and trained a scoring function to map both results into the same space. The scoring function consisted of a sigmoid over the confidence of the detectors trained on the same held out portion of the training set. As none of the disambiguation examples discussed here rely on the specific identity of the actors, we did not detect their identity. Instead, any sentence which contains names was automatically converted to one which contains arbitrary "person" labels.

The sentences in our corpus have either two or three interpretations. Each interpretation has one or more associated videos where the scene was shot from a different angle, carried out either by different actors, with different objects, or in different directions of motion. For each sentence-video pair, we performed a 1-out-of-2 or 1-out-of-3 classification task to determine which of the interpretations of the corresponding sentence best fits that video. Overall chance performance on our dataset is 49.04%, slightly lower than 50% due to the 1-out-of-3 classification examples.

The model presented here achieved an accuracy of 75.36% over the entire corpus averaged across all error categories. This demonstrates that the model is largely capable of capturing the underlying task and that similar compositional cross-modal models may do the same. For each of the 3 major ambiguity classes we had an accuracy of 84.26% for syntactic ambiguities, 72.28% for semantic ambiguities, and 64.44% for discourse ambiguities.

The most significant source of model failures are poor object detections. Objects are often rotated and presented at angles that are difficult to recognize. Certain object classes like the telescope

are much more difficult to recognize due to their small size and the fact that hands tend to largely occlude them. This accounts for the degraded performance of the semantic ambiguities relative to the syntactic ambiguities, as many more semantic ambiguities involved the telescope. Object detector performance is similarly responsible for the lower performance of the discourse ambiguities which relied much more on the accuracy of the person detector as many sentences involve only people interacting with each other without any additional objects. This degrades performance by removing a helpful constraint for inference, according to which people tend to be close to the objects they are manipulating. In addition, these sentences introduced more visual uncertainty as they often involved three actors.

The remaining errors are due to the event models. HMMs can fixate on short sequences of events which seem as if they are part of an action, but in fact are just noise or the prefix of another action. Ideally, one would want an event model which has a global view of the action, if an object went up from the beginning to the end of the video while a person was holding it, it's likely that the object was being picked up. The event models used here cannot enforce this constraint, they merely assert that the object was moving up for some number of frames; an event which can happen due to noise in the object detectors. Enforcing such local constraints instead of the global constraint of the motion of the object over the video makes joint tracking and event recognition tractable in the framework presented here but can lead to errors. Finding models which strike a better balance between local information and global constraints while maintaining tractable inference remains an area of future work.

8 Conclusion

We present a novel framework for studying ambiguous utterances expressed in a visual context. In particular, we formulate a new task for resolving structural ambiguities using visual signal. This is a fundamental task for humans, involving complex cognitive processing, and is a key challenge for language acquisition during childhood. We release a multimodal corpus that enables to address this task, as well as support further investigation of ambiguity related phenomena in visually grounded language processing. Finally, we present a unified approach for resolving ambiguous descriptions of videos, achieving good performance on our corpus.

While our current investigation focuses on structural *inference*, we intend to extend this line of work to *learning* scenarios, in which the agent has to deduce the meaning of words and sentences from structurally ambiguous input. Furthermore, our framework can be beneficial for image and video retrieval applications in which the query is expressed in natural language. Given an ambiguous query, our approach will enable matching and clustering the retrieved results according to the different query interpretations.

Acknowledgments

This material is based upon work supported by the Center for Brains, Minds, and Machines (CBMM), funded by NSF STC award CCF-1231216. SU was also supported by ERC Advanced Grant 269627 Digital Baby.

References

Kobus Barnard, Matthew Johnson, and David Forsyth. 2003. Word sense disambiguation with pictures. In *Proceedings of the HLT-NAACL 2003 workshop on Learning word meaning from non-linguistic data-Volume 6*, pages 1–5. Association for Computational Linguistics.

L. E. Baum, T. Petrie, G. Soules, and N. Weiss. 1970. A maximization technique occuring in the statistical analysis of probabilistic functions of Markov chains. *The Annals of Mathematical Statistics*, 41(1):164–171.

Elia Bruni, Gemma Boleda, Marco Baroni, and Nam-Khanh Tran. 2012. Distributional semantics in technicolor. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 136–145. Association for Computational Linguistics.

Moreno I Coco and Frank Keller. 2015. The interaction of visual and linguistic saliency during syntactic ambiguity resolution. *The Quarterly Journal of Experimental Psychology*, 68(1):46–74.

Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. 2010. Every picture tells a story: Generating sentences from images. In *Computer Vision–ECCV 2010*, pages 15–29. Springer.

Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. 2010. Object detection with discriminatively trained part-based models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(9):1627–1645.

Yunchao Gong, Liwei Wang, Micah Hodosh, Julia Hockenmaier, and Svetlana Lazebnik. 2014. Improving image-sentence embeddings using large weakly annotated photo collections. In *Computer Vision–ECCV 2014*, pages 529–545. Springer.

Andrej Karpathy and Li Fei-Fei. 2014. Deep visual-semantic alignments for generating image descriptions. *arXiv preprint arXiv:1412.2306*.

Evan Kidd and Judith Holler. 2009. Children's use of gesture to resolve lexical ambiguity. *Developmental Science*, 12(6):903–913.

Chen Kong, Dahua Lin, Mayank Bansal, Raquel Urtasun, and Sanja Fidler. 2014. What are you talking about? text-to-image coreference. In *Computer Vision and Pattern Recognition (CVPR)*, pages 3558–3565. IEEE.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. 2011. Hmdb: a large video database for human motion recognition. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2556–2563. IEEE.

G Kulkarni, V Premraj, S Dhar, Siming Li, Yejin Choi, AC Berg, and TL Berg. 2011. Baby talk: Understanding and generating simple image descriptions. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1601–1608. IEEE Computer Society.

Angeliki Lazaridou, Nghia The Pham, and Marco Baroni. 2015. Combining language and vision with a multimodal skip-gram model. *CoRR*, abs/1501.02598.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014*, pages 740–755. Springer.

Jingen Liu, Jiebo Luo, and Mubarak Shah. 2009. Recognizing realistic actions from videos in the wild. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1996–2003. IEEE.

Margaret Mitchell, Xufeng Han, Jesse Dodge, Alyssa Mensch, Amit Goyal, Alex Berg, Kota Yamaguchi, Tamara Berg, Karl Stratos, and Hal Daumé III. 2012. Midge: Generating image descriptions from computer vision detections. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 747–756. Association for Computational Linguistics.

Vignesh Ramanathan, Armand Joulin, Percy Liang, and Li Fei-Fei. 2014. Linking people in videos with their names using coreference resolution. In *Computer Vision–ECCV 2014*, pages 95–110. Springer.

Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. 2013. Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics*, 1:25–36.

Mikel D. Rodriguez, Javed Ahmed, and Mubarak Shah. 2008. Action MACH A Spatio-temporal Maximum Average Correlation Height Filter for Action Recognition. In *Computer Vision and Pattern Recognition*, pages 1–8.

Narayanaswamy Siddharth, Andrei Barbu, and Jeffrey Mark Siskind. 2014. Seeing what you're told: Sentence-guided activity recognition in video. In *Computer Vision and Pattern Recognition (CVPR)*, pages 732–739. IEEE.

Carina Silberer and Mirella Lapata. 2014. Learning grounded meaning representations with autoencoders. In *Proceedings of ACL*, pages 721–732.

Catherine E Snow. 1972. Mothers' speech to children learning language. *Child development*, pages 549–565.

Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. 2013. Zero-shot learning through cross-modal transfer. In *Advances in Neural Information Processing Systems*, pages 935–943.

Richard Socher, Andrej Karpathy, Quoc V Le, Christopher D Manning, and Andrew Y Ng. 2014. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics*, 2:207–218.

Michael J Spivey, Michael K Tanenhaus, Kathleen M Eberhard, and Julie C Sedivy. 2002. Eye movements and spoken language comprehension: Effects of visual context on syntactic ambiguity resolution. *Cognitive psychology*, 45(4):447–481.

Michael K Tanenhaus, Michael J Spivey-Knowlton, Kathleen M Eberhard, and Julie C Sedivy. 1995. Integration of visual and linguistic information in spoken language comprehension. *Science*, 268(5217):1632–1634.

Jesse Thomason, Subhashini Venugopalan, Sergio Guadarrama, Kate Saenko, and Raymond Mooney. 2014. Integrating language and vision to generate natural language descriptions of videos in the wild. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING), August*.

Subhashini Venugopalan, Huijuan Xu, Jeff Donahue, Marcus Rohrbach, Raymond Mooney, and Kate Saenko. 2015. Translating videos to natural language using deep recurrent neural networks. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Denver, Colorado.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Computer Vision and Pattern Recognition (CVPR)*.

A. J. Viterbi. 1971. Convolutional codes and their performance in communication systems. *Communications of the IEEE*, 19:751–772, October.