Inducing Structure for Vision and Language

Slav Petrov
Computer Science Division,
University of California, Berkeley
petrov@cs.berkeley.edu

A system that can observe events, learn, and participate just as a child would in an unknown scenario is the Holy Grail of AI.

1 Introduction

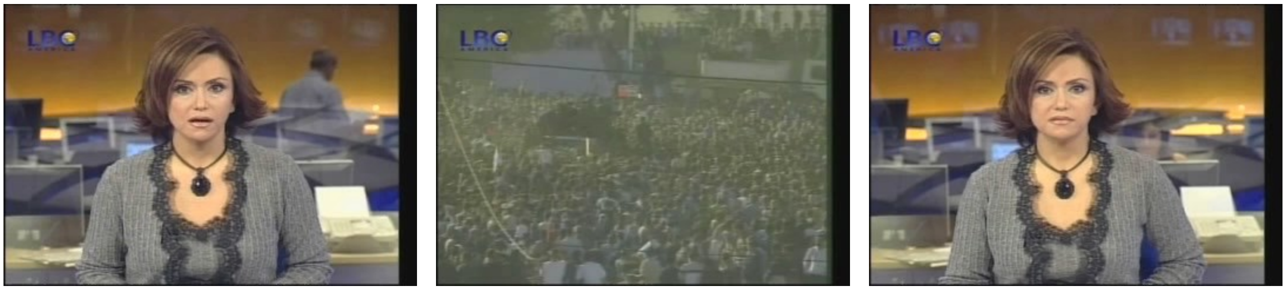
The ability of children to solve complex learning problems during their first years of life has fascinated philosophers and researchers throughout time. While we are still far from completely understanding this process, there has been some interesting recent work in learning and language grounding (Regier, 2003; Roy, 2005; Yu et al., 2005). The computational models presented therein are capable of discovering limited word and meaning associations from a video recording of a person telling a story shown in a picture book. Even though the models require a restricted and simplified environment, these results are remarkable if one considers the challenges involved: segmenting a speech signal into words, identifying objects and actions from their perceptual input, and associating these perceptions. Children are tackling these challenges with very little or no supervision when they acquire language. A computational model for learning and language grounding should therefore be able to (i) make use of multimodal information (e.g. visual and audio perception), (ii) be able to automatically induce models that capture the hidden structures present in language and vision, and (iii) do so in an unsupervised manner.

In the following I will discuss aspects of a computational model that meets the above requirements. First, I will briefly describe a system for learning connections between speech and images using unstructured models. While the system ranked among the best in a recent video retrieval challenge (NIST, 2006), its performance is clearly limited by the lack of structure in the model. I will contrast this to my work on learning the structure present in written text without detailed supervision. In this work, we developed a system that is capable of learning grammars for natural languages by automatically inducing latent structure models. These models achieve state of the art performance on parsing text written in English and a variety of other languages, clearly demonstrating the importance of structured models for perception. Finally, I will describe how I plan to apply techniques for unsupervised learning of language structure to visual data and form a unified computational model for learning language and vision.

2 Learning from Visual and Audio Information

Petrov et al. (2007) can be seen as a first step in the endeavor toward a system for learning from visual and audio information. We developed a system for object recognition and scene analysis in images and video with accompanying text streams (for example from web pages on the internet or TV shows). As an example application, we participated in the annual TRECVID challenge (NIST, 2006). The challenge requires the detection of 39 non-exclusive semantic categories ranging from objects (e.g., car, airplane), over scenes (e.g., mountain, waterfront) to activities (e.g., people marching) in broadcast news video. Figure 2 gives an example of three consecutive keyframes and the associated text which was output by an automatic speech recognizer (shortened and revised for illustration purposes).

Our approach considered three primary types of information present in the video signals. The raw audio signal was used to build a set of acoustic models that characterized the sounds associated with a given concept. More sophisticated automatic



(a) A crowd of tens of thousands Palestinians took farewell from its leader Yassir...

(b) ... Arafat, who was buried today ...

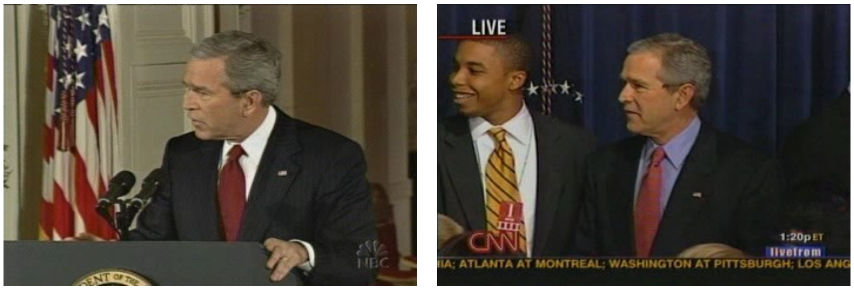
(c) ... after he passed away two days ago.

Figure 1: Typical example of temporal misalignment between visual and speech information. The second of the three consecutive shots shows a ”demonstration” while the most indicative words for ”demonstration” (e.g. *crowds*) are spoken during the first shot (text revised for illustration purposes). Unstructured models have difficulties modeling such relations.

speech recognition technology was deployed to decode the linguistic content of the audio signal; the decoded word sequences allowed us to treat this task as document classification. We also built a vision system that classified images from the video signal using shape descriptors. We then used unstructured bags of words/features models combined with a non-parametric classifier for detection.

Bag of features models offer only crude representations, but have been used in state of the art systems for object recognition (Zhang et al., 2006), and text retrieval. Even though we entered the challenge for the first time this year, our system ranked 6^{th} out of 27, ahead of many teams with several years of experience on this task. While our system performed reasonably well, a bag of words model is clearly not sufficient for this task. Consider the sequence shown in Figure 1 and assume we are trying to detect the category ”demonstration” using the speech transcripts only. The words that are most indicative for this category (e.g. *crowds*) are spoken in the studio and not while the demonstration is shown. It is not possible to convey this fact with a bag of words model, where all words are lumped together and the structural information is thrown away.

Another example of the necessity of structural information are the two images in Figure 2, of which the left image is a training image and the right image is a test image. At first sight, the two images appear very similar. Indeed, the bag of features representations were so similar that our system decided that the two images depict the same categories. This is correct for the categories ”person” and ”indoor scene” because they are shown in both the training and the test image. However, the test image does not contain a ”US-Flag”. This mistake can be attributed to the missing structural and semantic information that is not captured by the bag of features model. Structured models can use examples of objects shown in different contexts to learn which specific regions of the image correspond to the respective objects. Since this visual structure roughly corresponds to the grammar of a language, there is reason to believe that techniques for automatic refinement of natural language grammars could be effectively used also for learning the visual structure of scenes.



(a) Training Image

(b) Test Image

Figure 2: The test image in (b) is incorrectly assigned the label ”US-Flag” because the unstructured similarity to training image (a) is very large and the training image contains a ”US-Flag”. A structured model in cotrast can learn which parts of the image correspond to which category.

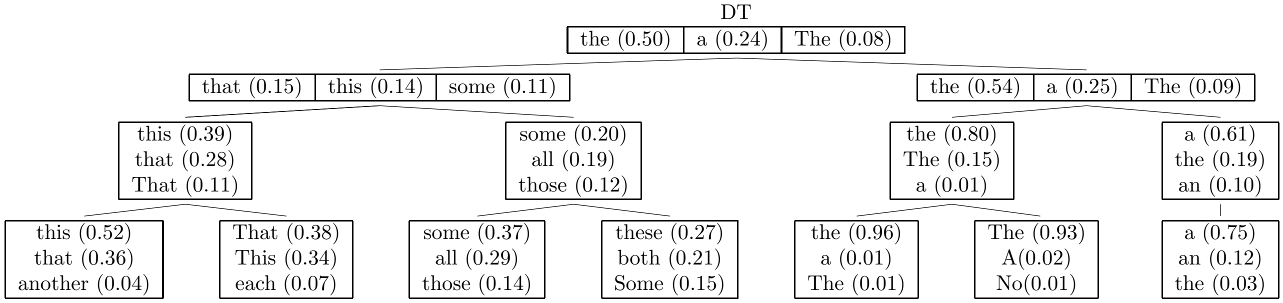

DT
the (0.50) a (0.24) The (0.08)
that (0.15) this (0.14) some (0.11)
the (0.54) a (0.25) The (0.09)
this (0.39) that (0.28) That (0.11)
some (0.20) all (0.19) those (0.12)
the (0.80) The (0.15) a (0.01)
a (0.61) the (0.19) an (0.10)
this (0.52) that (0.36) another (0.04)
That (0.38) This (0.34) each (0.07)
some (0.37) all (0.29) those (0.14)
these (0.27) both (0.21) Some (0.15)
the (0.96) a (0.01) The (0.01)
The (0.93) A(0.02) No(0.01)
a (0.75) an (0.12) the (0.03)

Figure 3: Evolution of the DT tag during hierarchical training. Shown are the top three words for each subcategory and their respective probability. Even though the refinements are induced in a completely unsupervised fashion, they form syntactically and semantically coherent groups.

3 Learning the Structure of Language

In the past year, we investigated data-driven approaches for learning language structure, encapsulated in the form of probabilistic context-free grammars (Petrov et al., 2006). We trained our models on a collection of hand parsed sentences (treebank), but without any additional human input. This is not trivial, since simply extracting the empirical rules and probabilities from a treebank results in a poor grammar (Charniak, 1996). The constituents of the treebank are not well suited for modeling language, because they imply unrealistic context-freedom assumptions. Therefore, a variety of techniques have been developed to both enrich and generalize the naive grammar by manually introducing annotations (Collins, 1999; Klein and Manning, 2003).

In our work we developed a simple, yet comprehensive framework for learning grammars with an automatic split-and-merge strategy. Beginning with the barest possible initial structure, we automatically refine each category, introducing more complexity only where needed. In general, any automatic induction system is in danger of being entirely uninterpretable. However, because of the way the latent structure is induced, our method reliably learns subcategories which exhibit most of the linguistically motivated splits discussed by previous work, e.g. in Klein and Manning (2003), and also captures other, additional linguistic phenomena.

As an example of how our staged training proceeds, Figure 3 shows the evolution of the subcategories of the determiner (DT) tag. In the first step demonstratives are split from determiners, then quantificational elements are split from demonstratives along one branch and definites from indefinites along the other, forming finer and finer syntactico-semantically coherent subcategories. Besides being automatically learned and linguistically interpretable, our models are very sparse. Our largest model is an order of magnitude smaller (300,000 vs. 3,000,000 parameters) than the previous state of the art model for parsing English (Charniak, 2000) and yet our model has a superior parsing accuracy.

4 Future Work

While structured models underlie most state of the art systems for natural language processing tasks (e.g. for word segmentation (Goldwater et al., 2006), parsing (Petrov et al., 2006) or summarization (Daumé III and Marcu, 2006)), vision systems typically resort to unstructured models (e.g. for object recognition (Zhang et al., 2006) or action recognition (Shechtman and Irani, 2005)). This can be partially explained by the fact that the structure of language is well defined in form of alphabet, words, and grammar, but the structure of vision is not.

It is my goal to address these hurdles and apply structural learning techniques for natural language to vision and, later, to form a unified model for learning language and vision. Figure 4 draws some parallels between visual and audio perception. While the correspondence is not meant to be exact, it clearly illustrates the similarities between vision and language. In the past, researchers have applied *unstructured* models from natural language processing to vision: object recognition has been treated as machine translation (Duygulu et al., 2002), or topic discovery methods have been employed for object detection and localization (Sivic et al., 2005). In contrast, I propose to apply *structured* models from natural language processing to vision. This distinction is a significant one, as recent work has shown the importance of structured models for vision (Sudderth et al.,

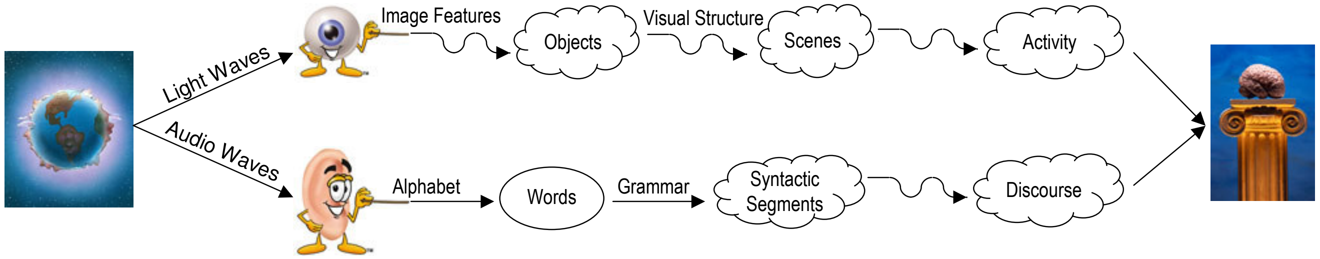


Figure 4: The main difference between visual and audio perception is that the structure of language is well defined in form of alphabet, words, and grammar, but the structure of vision is not.

2005; Lazebnik et al., 2006).

While this is an ambitious proposal, it has the advantage that proven techniques for structured learning in language can be brought to bear on vision tasks. Progress in the design of image features (Lowe, 1999; Berg et al., 2005) enables the design of a good ”visual alphabet” and my work on learning grammars provides a framework for automatically inducing a model of the visual structure of images. Joint learning of the vision and language components is facilitated by the parallel composition of the processing pipelines illustrated in Figure 4.

References

A. Berg, T. Berg, and J. Malik. 2005. Shape matching and object recognition using low distortion correspondence. In *CVPR ’05*.

E. Charniak. 1996. Tree-bank grammars. In *AAAI ’96*.

E. Charniak. 2000. A maximum–entropy–inspired parser. In *NAACL ’00*.

M. Collins. 1999. *Head-Driven Statistical Models for Natural Language Parsing*. Ph.D. thesis, U. of Pennsylvania.

H. Daumé III and D. Marcu. 2006. Bayesian query-focused summarization. In *ACL ’06*.

P. Duygulu, N. de Freitas, K. Barnard, and D. Forsyth. 2002. Object recognition as machine translation. In *ECCV ’02*.

S. Goldwater, T. Griffiths, and M. Johnson. 2006. Contextual dependencies in unsupervised word segmentation. In *ACL ’06*.

D. Klein and C. Manning. 2003. Accurate unlexicalized parsing. In *ACL ’03*.

S. Lazebnik, C. Schmid, and J. Ponce. 2006. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR ’06*.

D. Lowe. 1999. Object recognition from local scale-invariant features. In *ICCV ’99*.

NIST. 2006. Trec video retrieval evaluation 2001–2006. In *Digital Video Retrieval at NIST,* `http:\\www-nlpir.nist.gov/projects/trecvid`.

S. Petrov, L. Barrett, R. Thibaux, and D. Klein. 2006. Learning accurate, compact, and interpretable tree annotation. In *ACL ’06*.

S. Petrov, A. Faria, A. Berg, A. Stolcke, D. Klein, and J. Malik. 2007. Detecting categories in news video using acoustic, speech and image features. In *ICASSP ’07 (submitted)*.

T. Regier. 2003. Emergent constraints on word-learning: a computational perspective. In *Trends in Cognitive Science*.

D. Roy. 2005. Grounding words in perception and action: computational insights. In *Trends in Cognitive Science*.

E. Shechtman and M. Irani. 2005. Space-time behavioral correlation. In *CVPR ’05*.

J. Sivic, B. Russell, A. Efros, A. Zisserman, and B. Freeman. 2005. Discovering objects and their location in images. In *ICCV ’05*.

E. Sudderth, A. Torralba, W. Freeman, and A. Willsky. 2005. Learning hierarchical models of scenes, objects, and parts. In *ICCV ’05*.

C. Yu, D. Ballard, and R. Aslin. 2005. The role of embodied intention in early lexical acquisition. In *Cognitive Science*.

H. Zhang, A. Berg, M. Maire, and J. Malik. 2006. SVM-KNN: Discriminative nearest neighbor classification for visual category recognition. In *CVPR ’06*.