Toward Text-to-Picture Synthesis

Andrew B. Goldberg, Jake Rosin, Xiaojin Zhu, Charles R. Dyer
Department of Computer Sciences
University of Wisconsin–Madison
Madison, WI 53706
{goldberg,rosin,jerryzhu,dyer}@cs.wisc.edu

1 Introduction

It is estimated that more that 2 million people in the United States have significant communication impairments that result in them relying on methods other than natural speech alone for communication [2]. One type of commonly used augmentative and alternative communication (AAC) system is pictorial communication software such as SymWriter [8], which uses a lookup table to transliterate each word (or common phrase) in a sentence into an icon. This is an example of converting information between modalities. However, the resulting sequence of icons can be difficult to understand.

We have been developing general-purpose Text-to-Picture (TTP) synthesis algorithms [10, 5] to improve understandability using machine learning techniques. Our goal is to help users with special needs, such as the elderly or those with disabilities, to rapidly browse documents through pictorial summaries (e.g., Figure 5). Our TTP system targets general English. This differs from other pictorial conversion systems that require hand-crafted narrative descriptions of a scene [1, 9], 3D models [3], or special domains [6]. Instead, we use a concatenative or "collage" approach. In this talk, we discuss how machine learning enables the key components of our TTP system.

2 Extracting Text to Represent Pictorially

The first step in a concatenative approach to TTP synthesis is to identify the most salient pieces of text to draw. This is closely related to the natural language processing (NLP) tasks of information extraction and summarization, both of which have been the focus of much machine learning research. We have experimented with two approaches to this task.

Our first approach is keyword extraction with picturability [10]. The basic algorithm is a teleporting random walk (like PageRank) on a word graph called TextRank [7]. We modify the teleporting probabilities so that it prefers selecting words that are easy to visualize. Such "word picturability" is estimated from a logistic regression model trained on features derived from Web text and image search result counts. Figure 1 shows a set of words in this feature space, with symbols representing true classes.

Our second approach uses semantic role labeling (SRL)—the NLP task of deciding which words or phrases fulfill the various roles involved in a verb [5]. For example, the verb "to give" has roles for the person who does the giving, the object being given, and the person receiving this object.

3 Selecting Images to Represent Text

Given a set of extracted text phrases, the next task is to select images to represent them. One option is to issue a query for each phrase on an image search engine. Because current image search engines are not perfect, this approach typically requires reranking the search results (e.g., by clustering [10]). In TTP, there is additional information available: the context of the complete sentence or paragraph we are trying to visualize. We are exploring machine learning methods to rerank images by context.

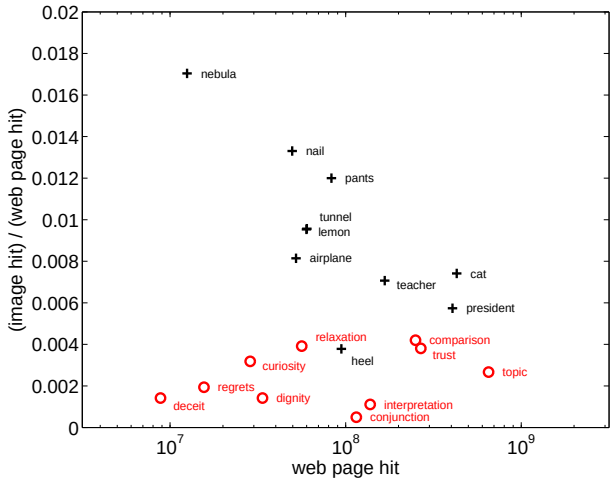


Figure 1: Visualizing word picturability—plus signs indicate concrete, picturable words, while circles indicate abstract, non-picturable concepts.

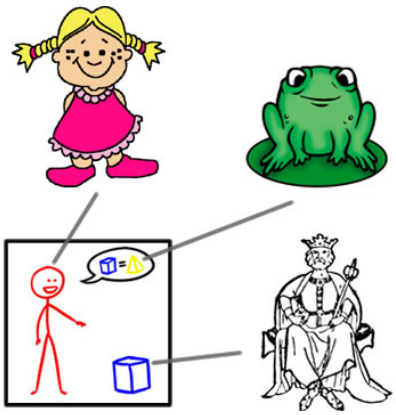


Figure 2: Illustration of the sentence "The girl called the king a frog" using a hand-drawn verb cartoon linked to noun images.

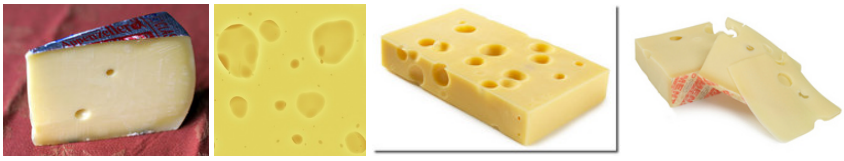


Figure 3: Using multiple images to represent a concept ("swiss cheese") may help to disambiguate it from a more general concept ("cheese").

Another facet of this problem for which machine learning can be useful is in determining how to select or produce images of nouns modified by adjectives, such as "black dog" or "fast dog."

Verbs have proven particularly challenging to visualize in our previous research. In our current work, we use hand-drawn cartoons illustrating common verbs (see Figure 2), with lines connecting each semantic role to its corresponding image. Since we cannot create cartoons for all verbs, we are investigating learning techniques to map a novel verb onto the closest verb for which we have a cartoon illustration. This learning task may involve both semantic and syntactic features to identify verbs that are interchangeable with respect to visual appearance.

Another issue is image sense ambiguity as perceived by the viewer. A high quality poodle image could be interpreted as "poodle," "dog," or "animal," following the hypernym relationship [4]. Our recent work has studied image sense disambiguation by presenting multiple images (Figure 3). For example, showing several poodle images at once will guide human perception onto "poodle." We have developed a Bayesian probabilistic model to explain such disambiguation.

4 Layout Optimization

After images have been selected, the final step requires that they be spatially arranged in a way that will help elicit the desired interpretation by users (i.e., help convey the original meaning of the text in question). This optimal layout can be learned in several ways. Our first TTP system [10] posed it as an optimization problem with an objective based on several criteria (Figure 4). Alternatively, layout optimization can be posed as a structured output prediction problem [5]. We designed a so-called ABC layout (Figure 5), such that each word (and ultimately its associated image) is tagged as being in the A, B, or C region using a linear-chain conditional random field.

In our current work, we are taking a more verb-centric approach using the cartoon drawings of verbs as in Figure 2. The layout problem reduces to deciding how to link images representing each

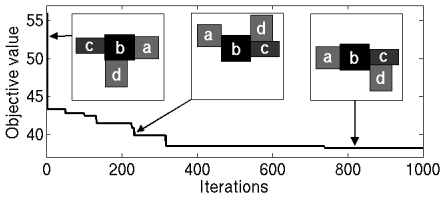
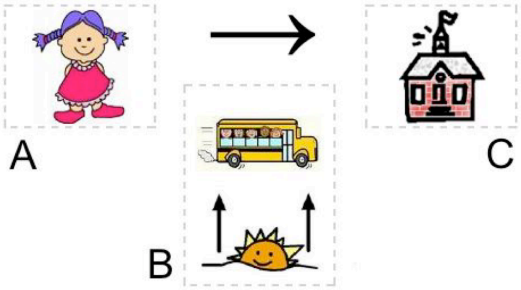


Figure 4: An example of picture layout optimization for images representing the text "a b c d." The objective function to be minimized measures image overlap, word importance, and the degree to which images representing words near each other in the text appear spatially close in the layout (image "a" should be closer to image "b" than "d").



The girl rides the bus to school in the morning
O A B B B O C O O B

Figure 5: Example ABC picture layout, original text, and tag sequence corresponding to the layout shown.

semantic role to the corresponding stick figures within the verb cartoon. We are experimenting with different linking visualization methods through user studies.

References

[1] G. Adorni, M. Di Manzo, and G. Ferrari. Natural language input for scene generation. In *ACL*, 1983.

[2] American Speech-Language-Hearing Association. Roles and responsibilities of speech-language pathologists with respect to augmentative and alternative communication: Technical report. *ASHA Supplement*, 24:1–17, 2004.

[3] B. Coyne and R. Sproat. WordsEye: An automatic text-to-scene conversion system. In *SIGGRAPH*, 2001.

[4] Christiane Fellbaum, editor. *Wordnet: An Electronic Lexical Database*. Bradford Books, 1998.

[5] A. B. Goldberg, X. Zhu, C. R. Dyer, M. Eldawy, and L. Heng. Easy as abc? facilitating pictorial communication via semantically enhanced layout. In *Proc. 12th Conf. Computational Natural Language Learning*, 2008.

[6] R. Johansson, A. Berglund, M. Danielsson, and P. Nugues. Automatic text-to-scene conversion in the traffic accident domain. In *IJCAI*, 2005.

[7] Rada Mihalcea and Paul Tarau. TextRank: Bringing order into texts. In *EMNLP'04*, 2004.

[8] Widgit Software. SymWriter, 2007. http://www.mayer-johnson.com.

[9] A. Yamada, T. Yamamoto, H. Ikeda, T. Nishida, and S. Doshita. Reconstructing spatial image from natural language texts. In *COLING*, 1992.

[10] X. Zhu, A. B. Goldberg, M. Eldawy, C. R. Dyer, and B. Strock. A Text-to-Picture synthesis system for augmenting communication. In *AAAI*, 2007.