Text-to-Picture Synthesis

Xiaojin Zhu

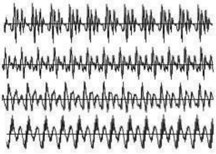
Department of Computer Sciences
University of Wisconsin–Madison

Joint work with Chuck Dyer, Andrew Goldberg,
Mohamed Eldawy, Bradley Strock, Lijie Heng, Art Glenberg

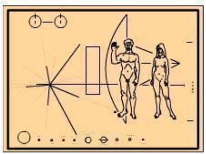
Outline

- prior work
- our first system
- our second system
- one application

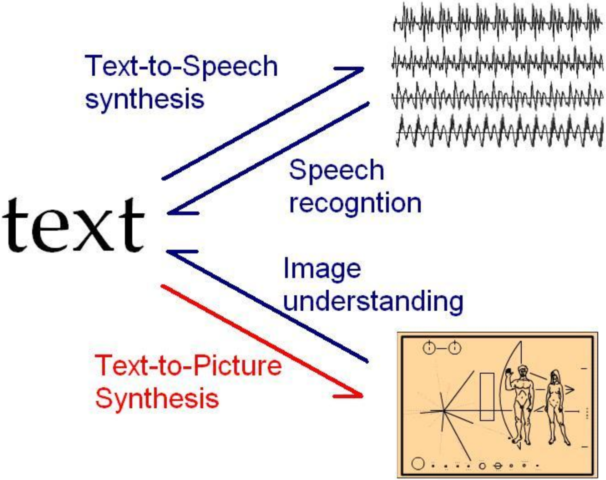
Humans switch modalities



text



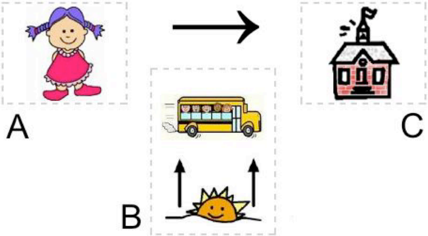
Computers switch modalities, too



Text-to-Picture (TTP) synthesis

Convert general natural language text into meaningful pictures.

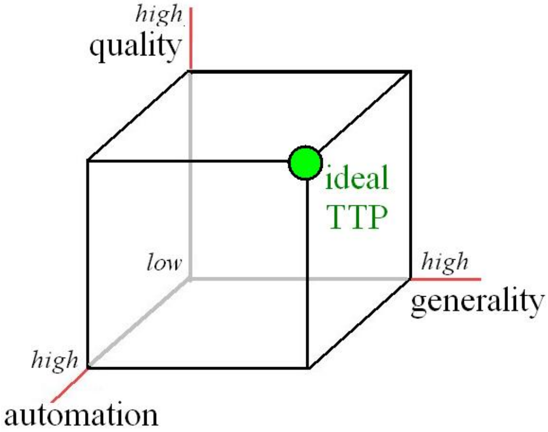
`The girl rides the bus to school in the morning.`



Applications of Text-to-Picture

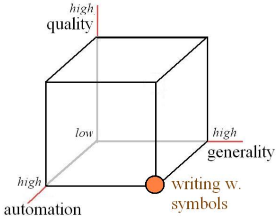
- Literacy development: young children, 2nd language speakers
- Assistive devices: people with learning disability
- Universal language
- Document summarization
- Image authoring tool

Three qualities of a Text-to-Picture system



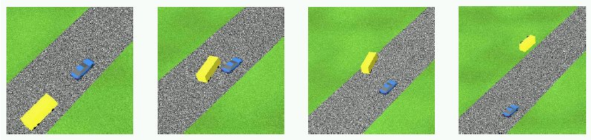
Prior work 1: “Writing with Symbols”

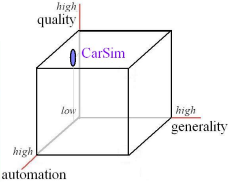
- Rebus symbols (`www.widgit.com`)
- Writing with Symbols (`www.mayer-johnson.com`)



Prior work 2: CarSim

A bus accident in southern Afghanistan last Thursday claimed 20 victims. Additionally, 39 people were injured in the accident, which occurred early Thursday morning twenty kilometers north of the city Kandahar. The bus was on its way from Kandahar towards the capital Kabul when it left the road while overtaking and overturned, said general Salim Khan, assistant head of police in Kandahar. The state of some of the injured was said to be critical.

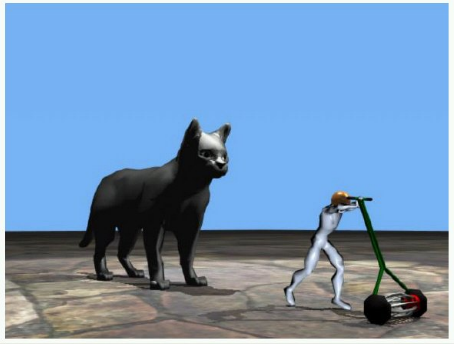


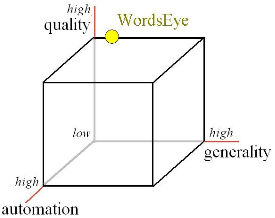


[Johansson, Berglund, Danielsson and Nugues. IJCAI 2005]

Prior work 3: WordsEye

`The lawn mower is 5 feet tall.` `John pushes the lawn mower.`
`The cat is 5 feet behind John.` `The cat is 10 feet tall.`

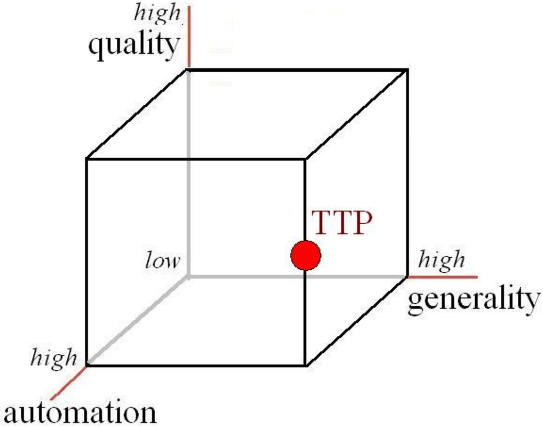




[Coyne and Sproat. SIGGRAPH 2001] (`www.wordseye.com`)

Our TTP system

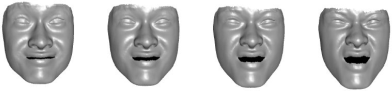
Our TTP system should handle general text automatically.



Approaches to Text-to-Picture

1. "Canned" pictures

2. Model-based



3. Concatenative (our system)

```
First the farmer gives
hay to the goat.  Then
the farmer gets milk from
the cow.
```

→

Components of our TTP systems

1. Keyphrase selection
2. Image selection
3. Layout
4. Evaluation

Our first TTP system

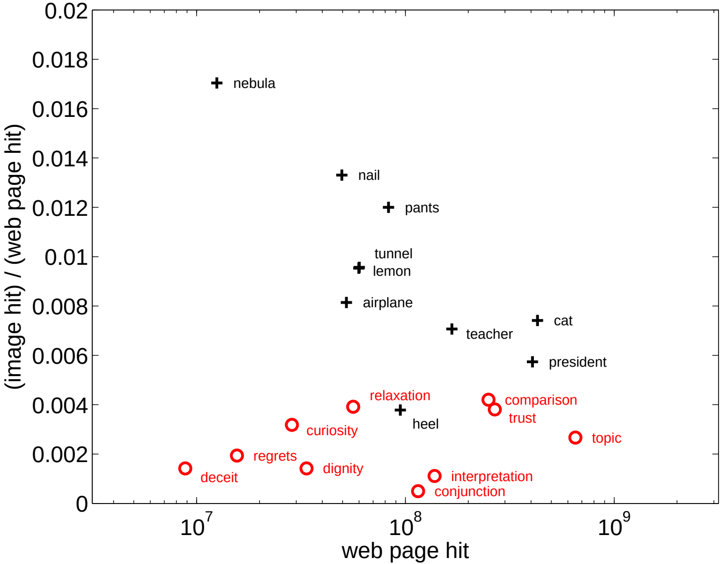
Step 1: Keyphrase selection

Problem: decide which words to draw.

- TextRank keyword summarization [Mihalcea and Tarau 2004].
- Graph on nouns, proper nouns, adjectives.
- Edges for word co-occurrence.
- Random walk on the graph.
- Stationary distribution (PageRank) as word importance.
- Important difference: word picturability used for teleporting probability.

Word picturability

We want to select picturable words.



Word picturability model

- Labels provided by five human annotators
  0 1 0 0 0 writ
  1 1 1 1 1 yolks
  1 1 1 1 1 zebras
  1 0 1 0 1 zigzag
- 253 candidate features from Google, Yahoo!, Flickr search
- Best feature: $x = \log(\text{Google image hits}/\text{Google Web page hits})$
- Logistic regression with forward feature selection
- $\Pr(\text{picturable}|x) = 1/\left(1 + \exp(-2.78x - 15.4)\right)$

Step 2: Image selection

Problem: find the best image for a word.

- Collect top image search results

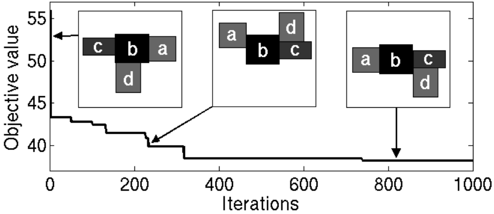
- Segmentation
- Cluster image segments by color
- Select the image containing the segment at the center of the largest cluster.

Step 3: Image layout

Problem: put the images together.

- minimum overlap
- important words at center
- close in text, close in picture

Stochastic optimization.

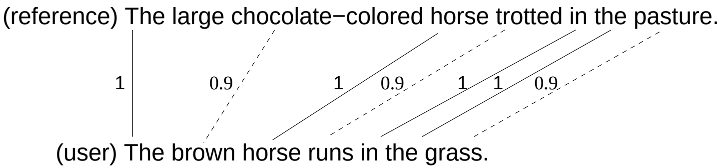


Step 4: Evaluation

The large chocolate-colored horse trotted in the pasture.

The brown horse runs in the grass.

Evaluation

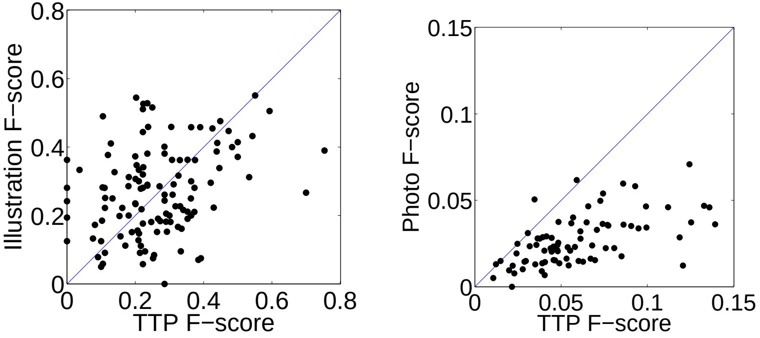


- synonyms
- greedy word alignment
- F-measure from precision and recall

User study

story text	original illustration
↓	
TTP picture	↓
↓	
user text (TTP)	user text (illustration)
↓	↓
F-score (TTP)	F-score (illustration)

User study results



TTP vs. original illustration	News photo vs. TTP + photo
6 users, 40 stories	8 users, 20 stories
(children's story)	(news)

Our first system is far from perfect

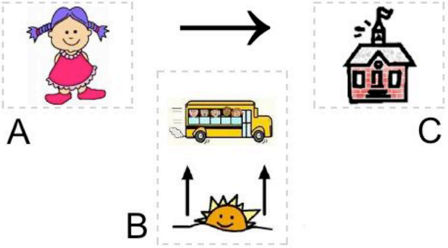
```
The girl loved the dog.  The
girl loved the dog's soft
eyes and warm nose and big
paws.  The girl wished she
had a dog.
```

"A girl's pet puts its paw on her nose."
"The dog walked up to the girl and sniffed her."
"The dog bit the girl in her nose and ran away."
"The girl's nose smelled the dog and monkey as they walked away."
"The girl walked her dog and saw a hairy man with a big nose."
"The girl monkey nose smells dog paw prints."

Our second TTP system

ABC layout

- Inspired by pilot user study
- 3 positions and an arrow
- Positions ≈ semantic roles
  - A = "who"
  - B = "what action" / "when"
  - C = "to whom" / "for what"
- Function words omitted

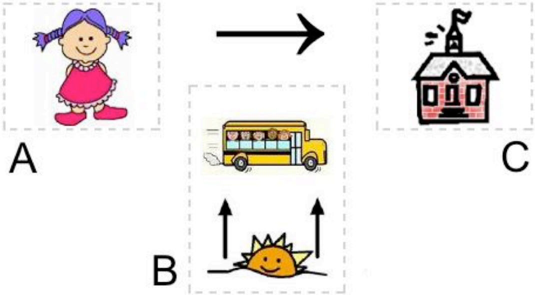


Advantages

- Structure helps disambiguate icons (verb vs. noun)
- Learnable by casting as a sequence tagging problem

ABC layout prediction as sequence tagging

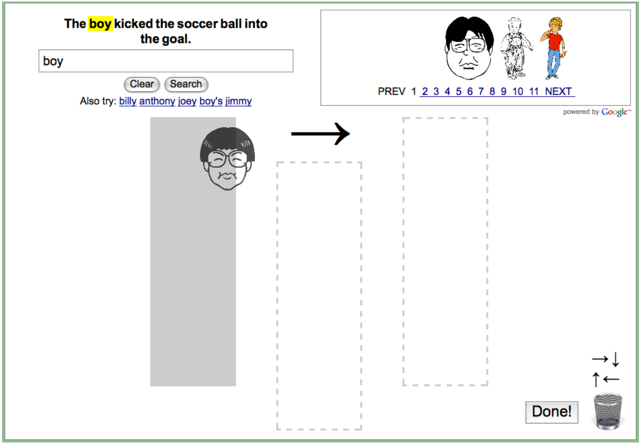
Given input sentence, assign {A, B, C, O} tags to words



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

Obtaining training data for layout predictor

Web-based "pictionary"-like tool to create ABC layouts for
571 sentences from school texts, children's books, news headlines
For 48 texts, 3 annotators: tag agreement = 77%, Fleiss' kappa = 0.71



Chunking by Semantic Role Labeling

Note: We actually work at chunk level; word level is too fine-grained.

Obtain semantically coherent chunks as basic units in the pictures

- Assign PropBank semantic roles using ASSERT [Pradhan et al. 2004]
- We use SRL *as is*—used model provided with ASSERT
- PropBank roles define chunks to be placed in layout

Example:

The boy	gave	the ball	to	the girl	yesterday
↑	↑	↑		↑	↑
Arg0	Target	Arg1		Arg2	ArgM-TMP

Sequence tagging with linear-chain CRFs

Goal: Tag each chunk with a label in $\{$A,B,C,O$\}$

Input: Chunk sequence $\mathbf{x}$ and features

Output: Most likely tag sequence $\mathbf{y}$

$\mathbf{y} =$	**A**	**B**	**B**		**C**	**B**
$\mathbf{x} =$	The boy	gave	the ball	to	the girl	yesterday
	$\uparrow$	$\uparrow$	$\uparrow$		$\uparrow$	$\uparrow$
	Arg0	Target	Arg1		Arg2	ArgM-TMP

Note: Each chunk described by PropBank and other features

Sequence tagging with linear-chain CRFs

Probabilistic model:

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \exp\left(\sum_{t=1}^{|\mathbf{x}|} \sum_{k=1}^{K} \lambda_k f_k(y_t, y_{t-1}, \mathbf{x}, t)\right),$$

Different factorizations of $\lambda_k f_k(y_t, y_{t-1}, \mathbf{x}, t)$:

- Model 1: Tag sequence ignored; one weight for each tag-feature
- Model 2: HMM-like; weights for transitions and emissions
- Model 3: General linear-chain; one weight per tag-tag-feature

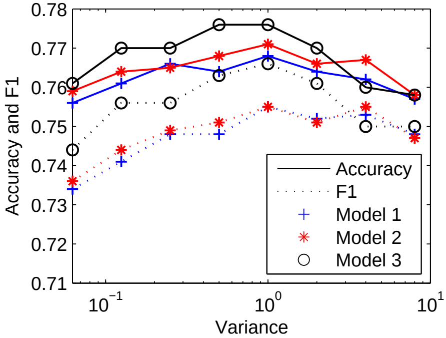
CRF Features

Binary predicate features evaluated for each SRL chunk

1. PropBank role label of the chunk
   - e.g., Arg0? Arg1? ArgM-LOC?
2. Part-of-speech tags of all words in the chunk
   - e.g., Contains JJ? NNP? RB?
3. Features related to the type of phrase containing the chunk
   - e.g., NP? PP? Is the chunk inside a VP?
4. Lexical features: 5000 frequent words and WordNet supersenses
   - e.g., Contains 'girl'? 'pizza'? verb.consumption?

CRF Experimental Results

To choose model and CRF's regularization parameter, ran 5-fold cross validation



Best accuracy and macro-avg F1 achieved with Model 3, $\sigma^2 = 1.0$

Accuracy is similar to that of human annotators

User Study: Is ABC layout more useful than linear layout?

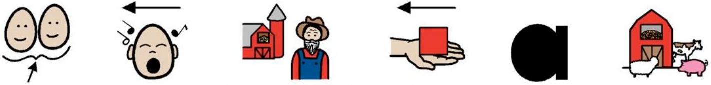
Subjects: 7 non-native English speakers, 12 native speakers
90 test sentences from important TTP application domains
Each subject saw 45 linear pictures and 45 ABC pictures

User study overall protocol

original sentence
↓
SymWriter icons
↙ ↘

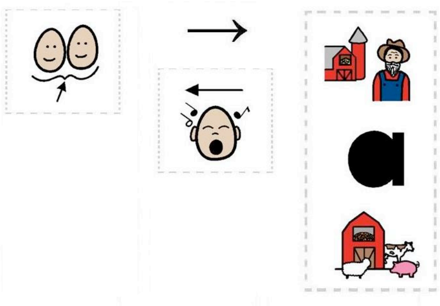
ABC layout	Linear layout
↓	↓
user text	user text
↓	↓
BLEU/METEOR (ABC)	BLEU/METEOR (Linear)

Sample picture and guesses: Linear layout



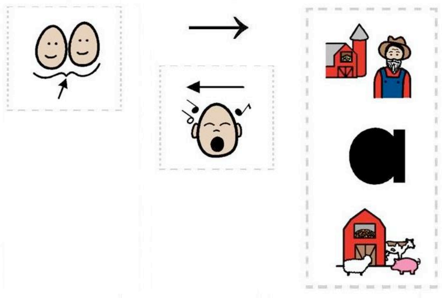
"we sing a song about a farm."
"i sing about the farm and animals"
"we sang for the farmer and he gave us animals."
"i can't sing in the choir because i have to tend to the animals."

Sample picture and guesses: ABC layout



"they sing old mcdonald had a farm."
"we have a farm with a sheep, a pig and a cow."
"two people sing old mcdonald had a farm"
"we sang old mcdonald on the farm."

Sample picture and guesses: ABC layout



"they sing old mcdonald had a farm."
"we have a farm with a sheep, a pig and a cow."
"two people sing old mcdonald had a farm"
"we sang old mcdonald on the farm."

Original: We sang Old MacDonald had a farm.

Results of user study

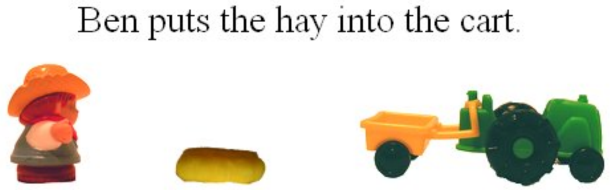
	Non-native		Native	
	ABC	Linear	ABC	Linear
METEOR	**0.1975**	0.1800	0.2955	**0.3335**
BLEU	**0.1497**	0.1456	0.2710	**0.3011**
Time	**47.4s**	47.8s	**38.1s**	38.6s

- ABC layout allows non-native speakers to recover more meaning
- However, the linear layout is better for native speakers
  - Familiar with left-to-right structure of English
  - Can guess the meaning of obscure function-word icons
- More complex layout does not require additional processing time

One application:
Improving children's reading comprehension

Physical activity can enhance young children's reading comprehension

- The Indexical Hypothesis (Glenberg, Gutierrez and Levin 2004)
  - Young readers may not "index" (map) words to objects
  - Consequently, they fail to derive meaning from text
  - New instructional method: manipulating toys according to text



- Physical manipulation results in better memory for and comprehension of the text.
- But: pain of real toys

Computer manipulation

- Will manipulating images on a computer have the same effect as manipulating physical toys?
- Computer images generated manually (TTP in the future).
- 53 1st and 2nd grade children.
- Three conditions:
  - physical manipulation (PM)
  - computer manipulation (CM)
  - re-read without manipulation (CR)
- Memory/comprehension test after each story.

Example story

At the barn

Ben puts the hay into the cart.

He drives the cart into the barn.

He needs to put the hay away.

Later, he will use the hay to feed the animals.

Ben puts the hay into the hayloft

http://www.cs.wisc.edu/zhu/space2/psych/new/interface.php

CM $\geq$ PM $\geq$ CR

- Measure: proportion correct for the memory/comprehension test questions

Condition	N	Correct
CM	20	$.89 \pm .06$
PM	14	$.84 \pm .11$
CR	19	$.80 \pm .10$

- CM>CR significant at $p = 0.01$
- CM as good as PM: opens up many doors

Conclusions

1. Text-to-Picture is an interesting and complex research topic.
2. We have some preliminary ideas, much still needs to be done.

Funding acknowledgment: NSF IIS-0711887, Wisconsin Alumni Research Foundation.

Backup Slides

Why not use manual rules from PropBank to ABC?

PropBank roles are verb-specific

- Arg0 is typically the agent, but Arg1, Arg2, etc. do not generalize
- For example, Arg1 can map to either B or C:

Bob$_{Arg0}$	$\longrightarrow$	Sue$_{Arg2}$
	gave$_{Target}$ book$_{Arg1}$	

Bob$_{Arg0}$	$\longrightarrow$	car$_{Arg1}$
	drove$_{Target}$	

Other issues

- Best position of modifiers like ArgM-LOC depends on usage
- Sentences with multiple verbs need special treatment

Bottom line

Mapping from semantic roles to layout positions is non-trivial!

CRF Experimental Results

Relative importance of the types of features

- Lexical > PropBank labels > phrase tags > part-of-speech tags

Learned feature weights make intuitive sense

- Preferred tag transitions: A $\rightarrow$ B, B $\rightarrow$ C
- Preferred in A: noun phrases (not nested in verb phrase)
- Preferred in B: verbs and ArgM-NEGs
- Preferred in C: supersense noun.objects, Arg4s, and ArgM-CAUs

Error analysis reveals similar mistakes as human annotators. Accuracy is similar to inter-annotator agreement.

Conclusion

The CRF model *can* predict the layouts about as well as humans.