Easy as ABC? Facilitating Pictorial Communication via Semantically Enhanced Layout

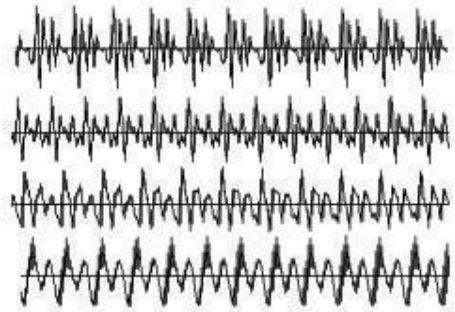
Andrew B. Goldberg Xiaojin Zhu Charles Dyer
Mohamed Eldawy Lijie Heng

Department of Computer Sciences
University of Wisconsin, Madison, USA

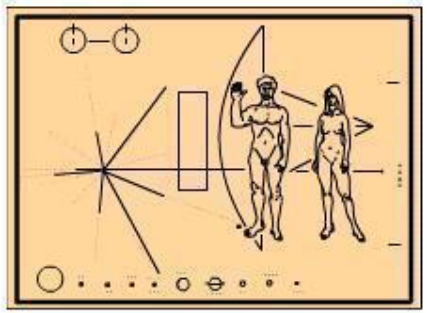
Presented by Kenji Sagae, USC Institute for Creative Technologies

CoNLL 2008

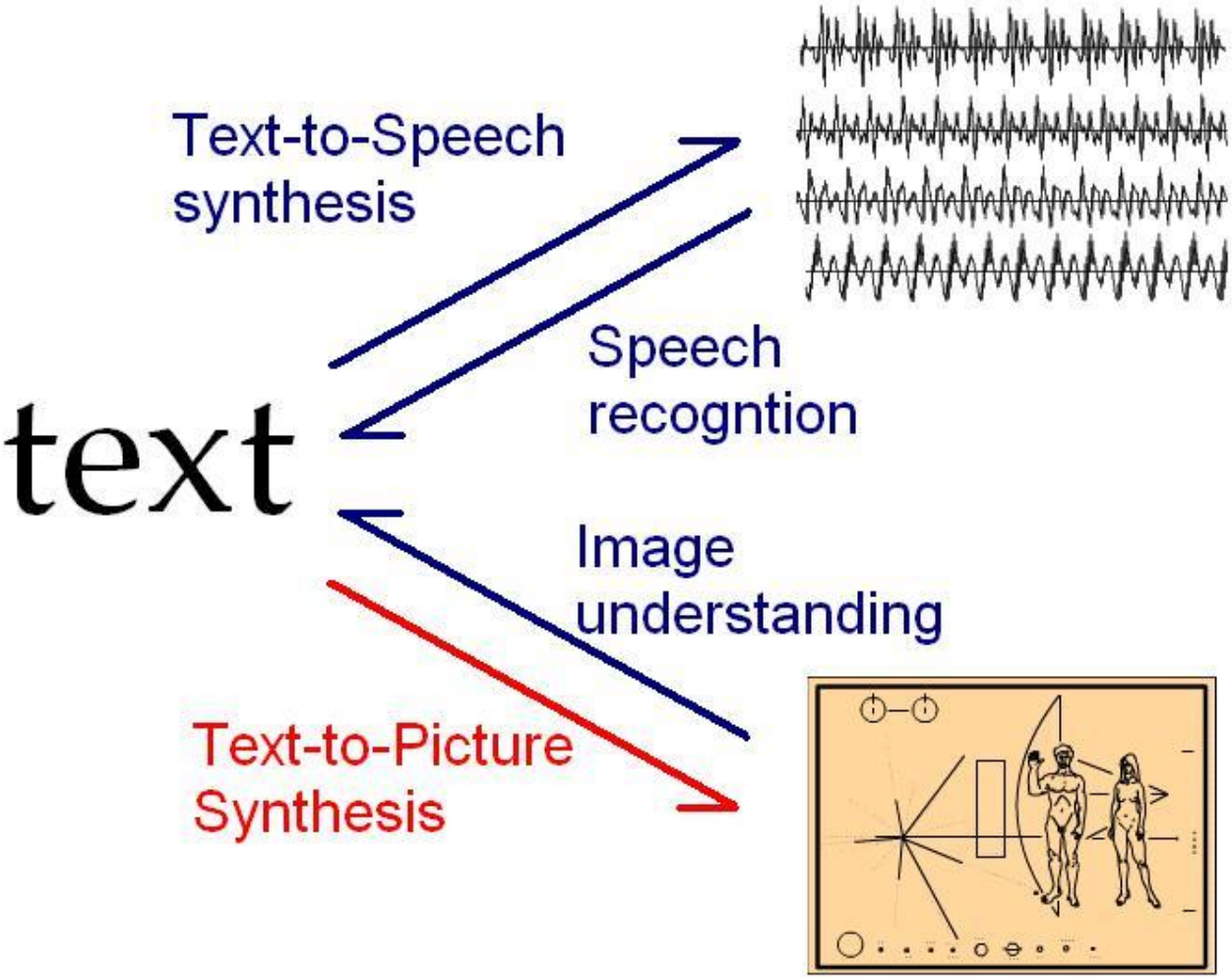
Humans communicate in multiple modalities



text

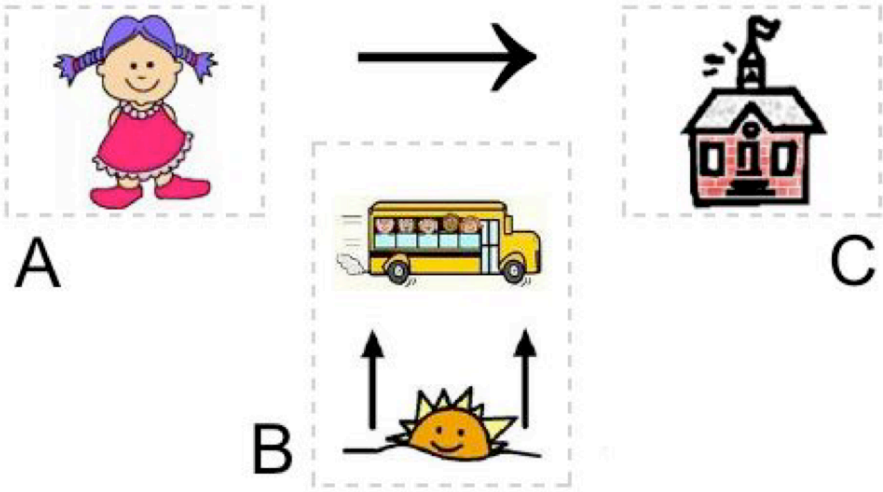


Computer modalities



Text-to-Picture (TTP) synthesis (aka Pictorial Communication)

The girl rides the bus to school in the morning.



Goal

Convert general natural language text into meaningful pictures for:

- Literacy development: young children, 2nd language speakers
- Assistive devices: people with learning disabilities
- Universal language, document summarization, image authoring tool

Outline

1. The picture layout problem
2. Predicting layouts using semantic role labeling, syntactic parsing, and conditional random fields
3. User study results

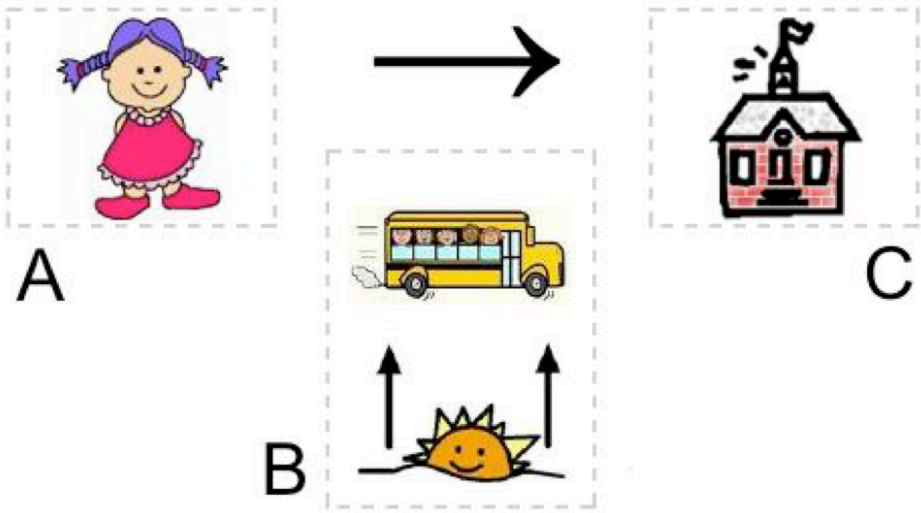
Components of our TTP system [Zhu et al. AAAI 2007]

"Collage" approach involving three main steps:

1. Keyphrase selection
2. Image selection
3. Picture layout:
   - Given an input sentence and set of icons
   - Produce layout that best conveys the meaning of the input text
   - Current work: Predict novel "ABC" layout using CRFs.

ABC layout

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

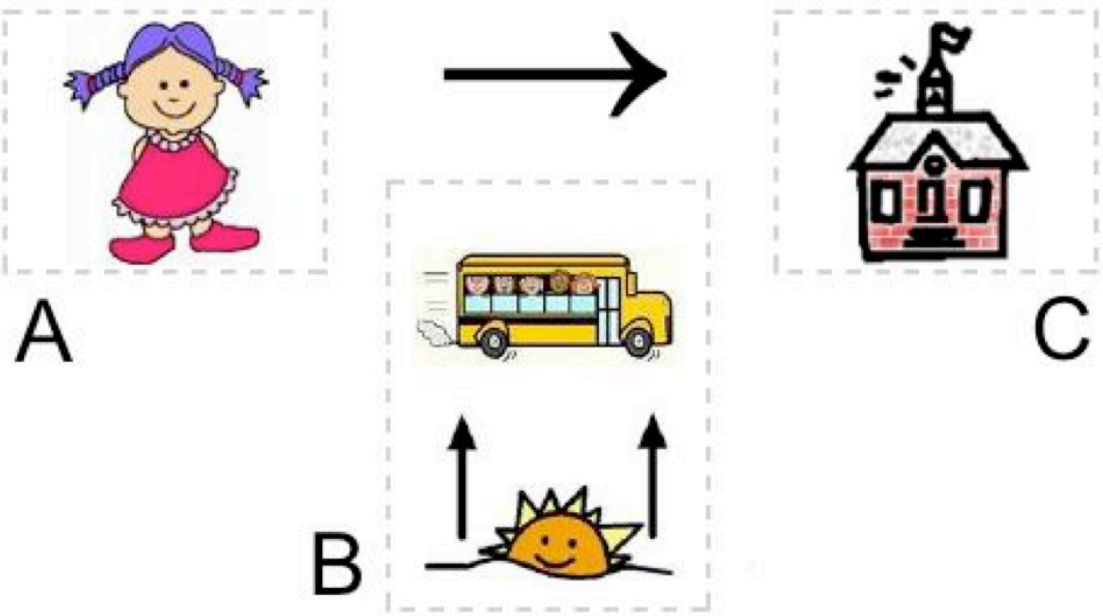


Advantages

- 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



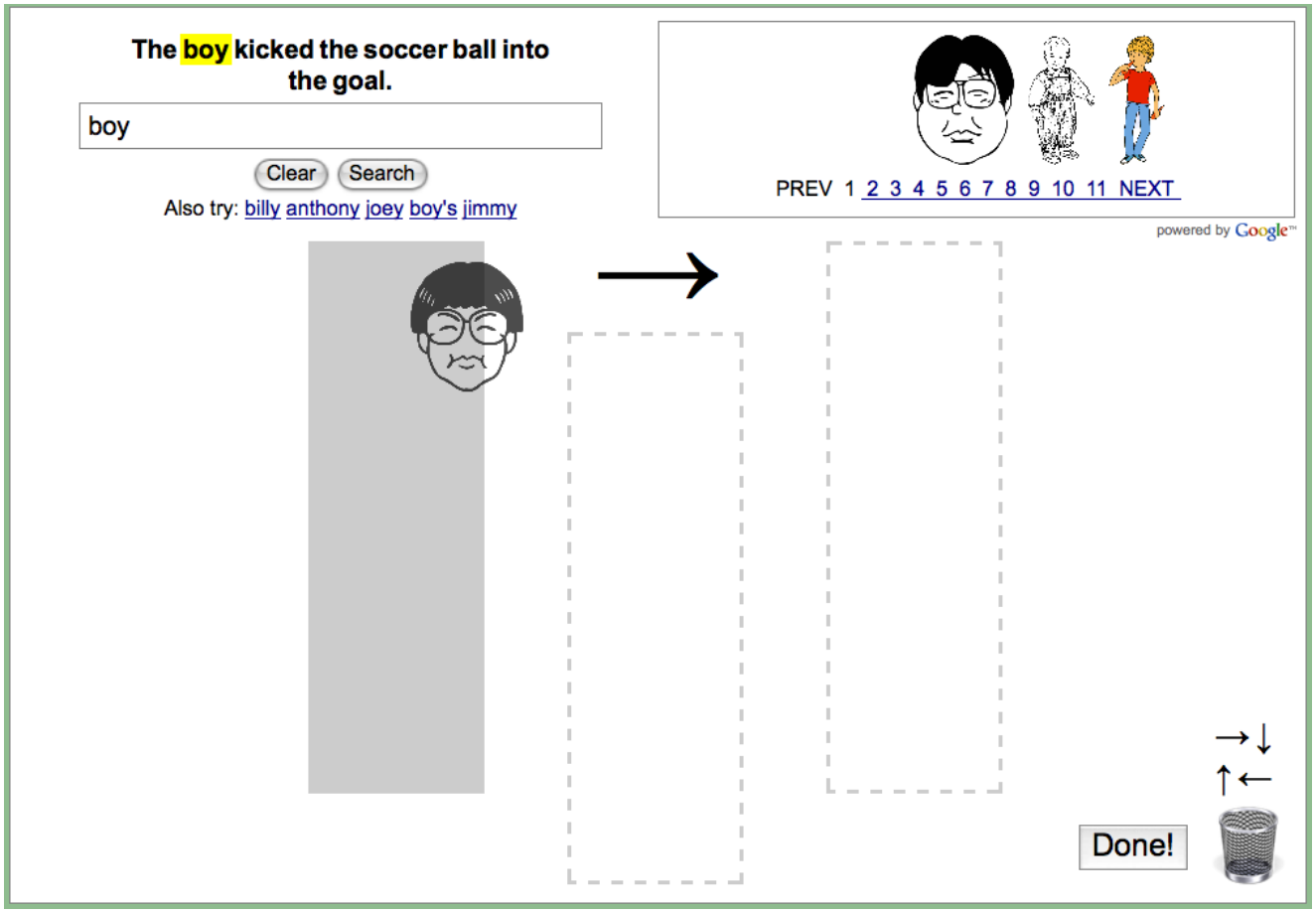
The girl rides the bus to school in the morning

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



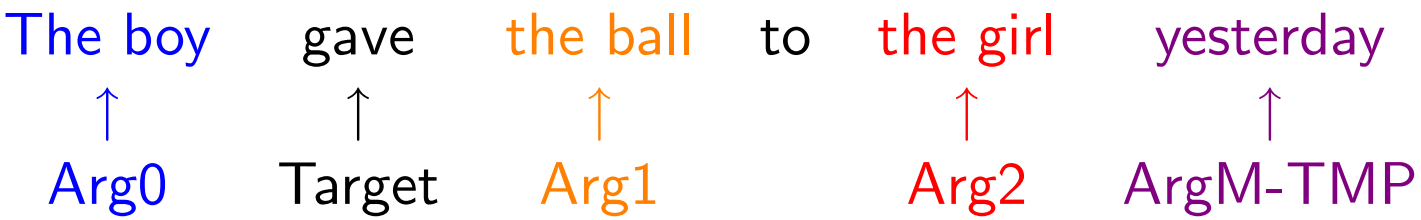
Chunking by Semantic Role Labeling

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

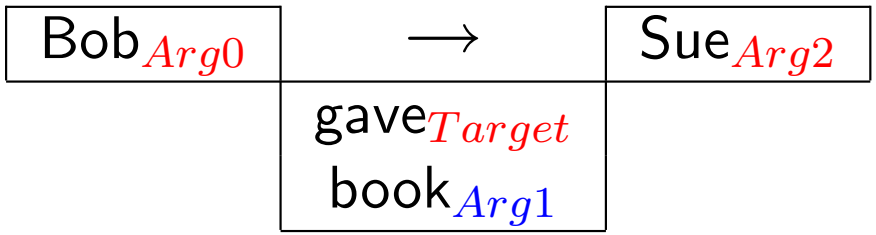
Example:

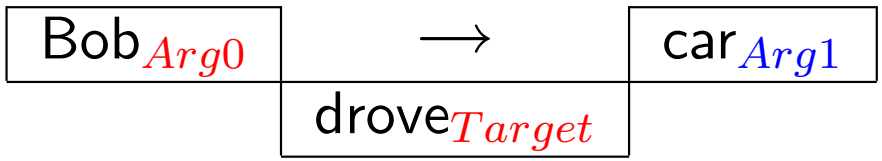


Why not use manual rules from PropBank to ABC?

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:





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!

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

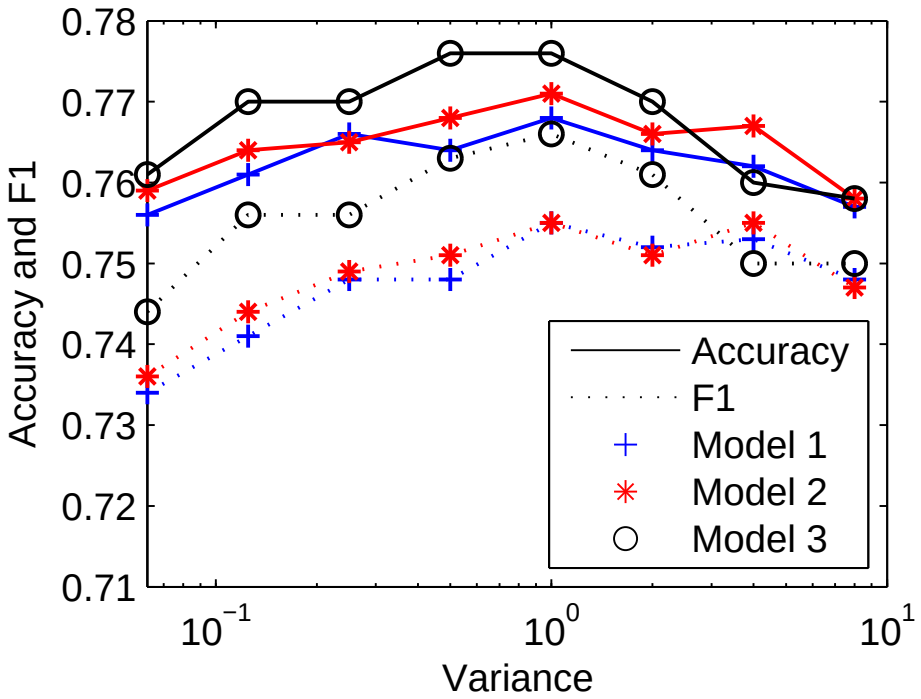
CRF Features

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

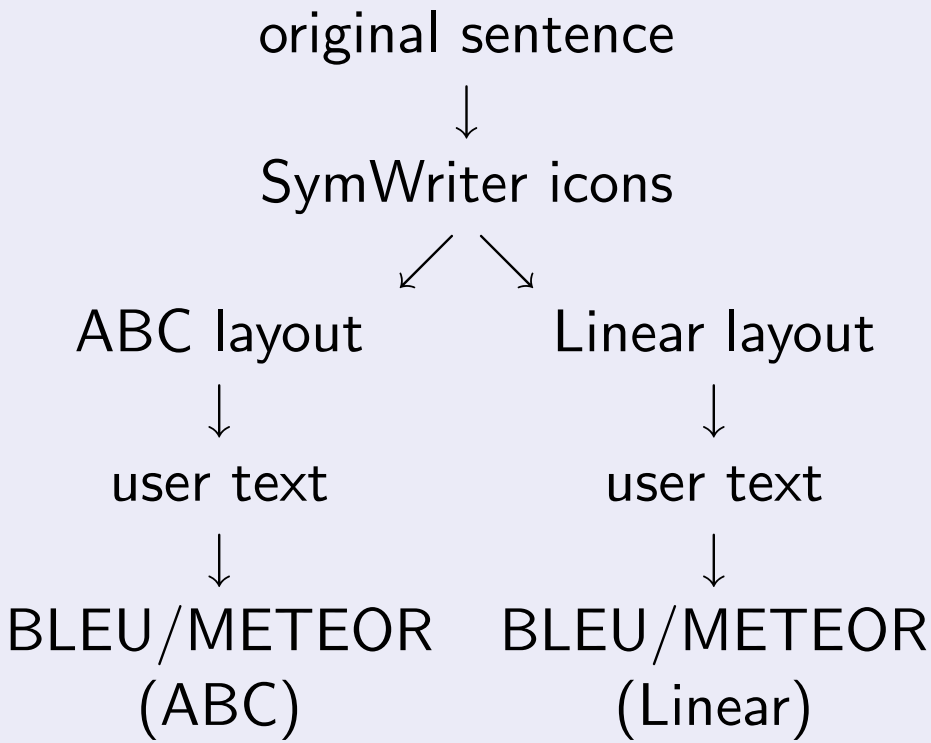


Best accuracy and macro-avg F1 achieved with Model 3, $\sigma^2 = 1.0$
Accuracy is similar to that of human annotators

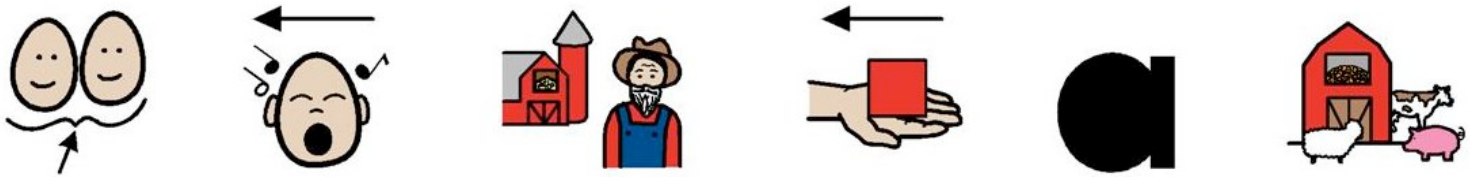
User Study: Is ABC layout more useful than linear layout?

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

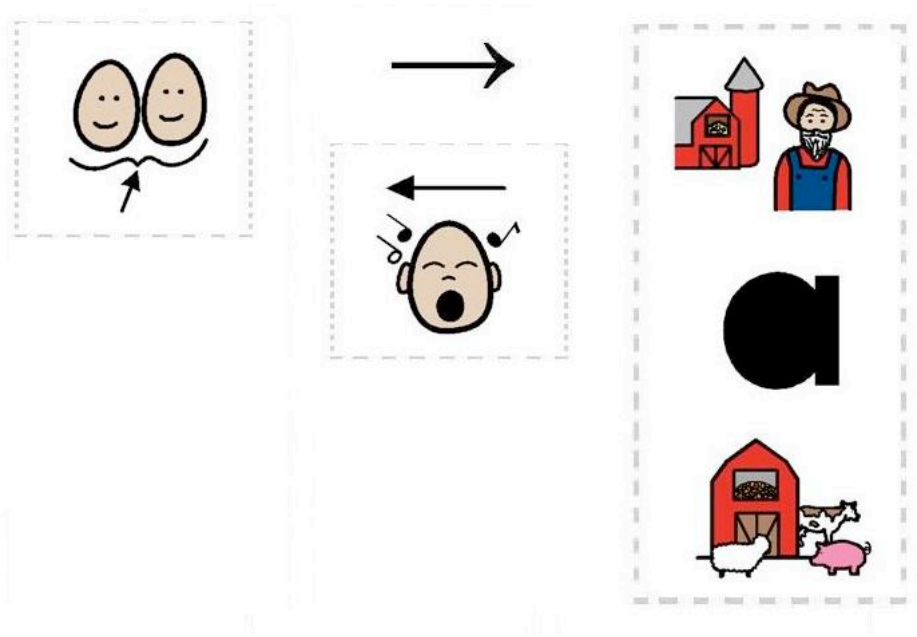


Sample picture and guesses: Linear layout



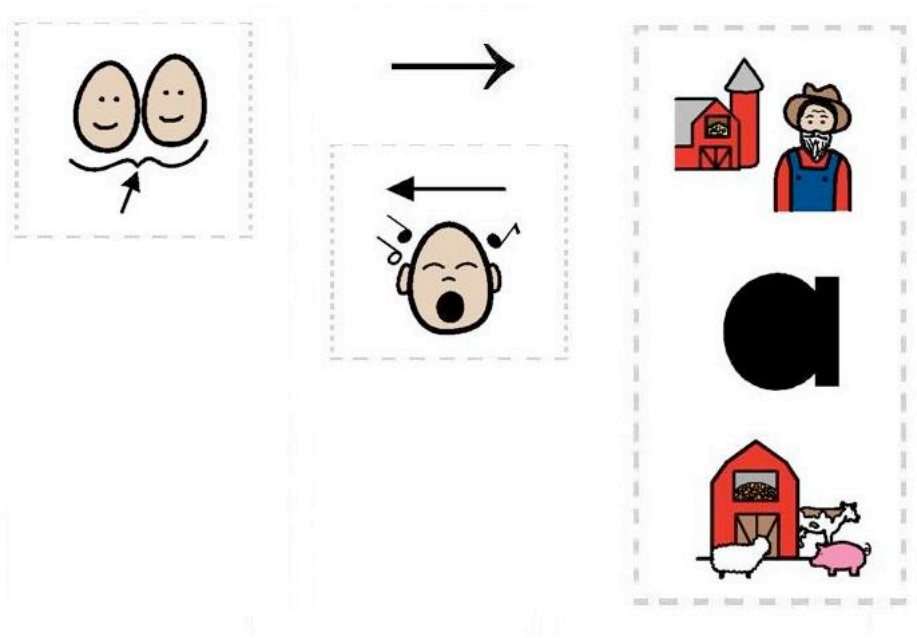
"we sing a song about a farm."
"i sing about the farm and animals"
"we sang for the farmer and he gave us animals."
"i can't sing in the choir because i have to tend to the animals."

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

Conclusions

1. Proposed a semantically enhanced picture layout for pictorial communication
2. Formulated our ABC layout prediction problem as sequence tagging
3. Leveraged semantic role labeling to segment text into picture units
4. Trained CRF layout prediction models with linguistic features
5. User study suggests ABC layout has potential to help picture comprehension in people with limited English literacy

Future work:

- Incorporate ABC layouts in our larger TTP system
- Use NLP and computer vision techniques to select icon(s) for each semantic chunk

Conclusions

1. Proposed a semantically enhanced picture layout for pictorial communication
2. Formulated our ABC layout prediction problem as sequence tagging
3. Leveraged semantic role labeling to segment text into picture units
4. Trained CRF layout prediction models with linguistic features
5. User study suggests ABC layout has potential to help picture comprehension in people with limited English literacy

Future work:

- Incorporate ABC layouts in our larger TTP system
- Use NLP and computer vision techniques to select icon(s) for each semantic chunk

Thank you.

Backup Slides

Representative prior work

"Writing with Symbols" [SymWriter (`www.mayer-johnson.com`)]

- "Transliterates" words into icons one at a time
- Little human effort, but requires familiarity with symbol set

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

- Specialized system creates images based on car-accident descriptions

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

- Creates 3D scenes based on scene descriptive language

Goal of our overall project

To convey the gist of general, unrestricted text.

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.