Generating bilingual pragmatic color references

Will Monroe[1] Jennifer Hu[2] Andrew Jong[3] Christopher Potts[1]

[1]Stanford University, [2]Harvard University, [3]San José State University

Main contribution

- A **bilingually trained** model
- for a **grounded language generation** task
- can learn **abstract pragmatic behaviors** that hold across both languages
- while remaining sensitive to **language-specific differences**.

Task and dataset

A reference game: speaker has to describe a color to a listener, listener has to identify it among distractors.

We expand the dataset of Monroe et al. (2017) by collecting reference game utterances in Chinese. Example games:

鲜绿 *xiān lǜ*
[close] ‘bright green’

不亮的橙色 *bu-liàng de chéngsè*
[split] ‘not-bright orange’

紫红色 *zǐ hóngsè*
[far] ‘purple-red’

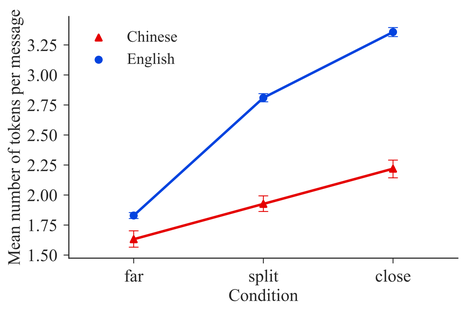
	en	zh
# pairs	948	117
# games	46,994	4,996
# utterances	53,365	6,534

Utterance statistics

English and Chinese display similar patterns of sensitivity to the difficulty of the context.

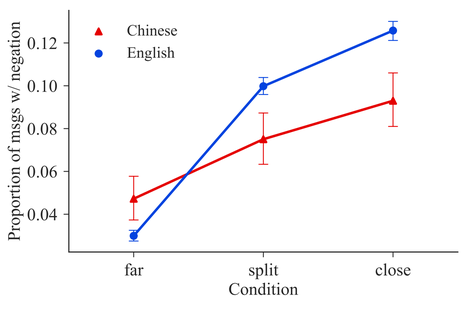
Length

tokens



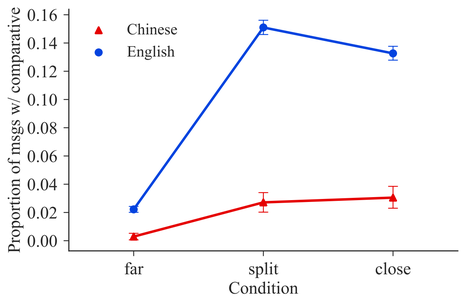
Negations

en: *not*, *n't*
zh: 不 *bù*, 没 *méi*



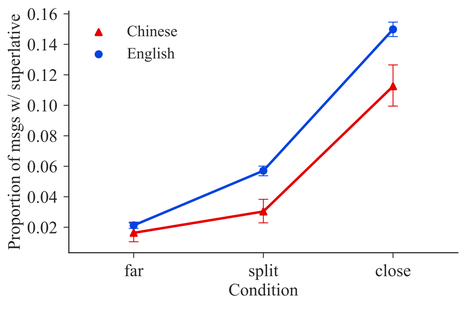
Comparatives

en: JJR, RBR
zh: 更 *gèng*, 比 *bǐ*



Superlatives

en: JJS, RBS
zh: 最 *zuì*



Get the dataset

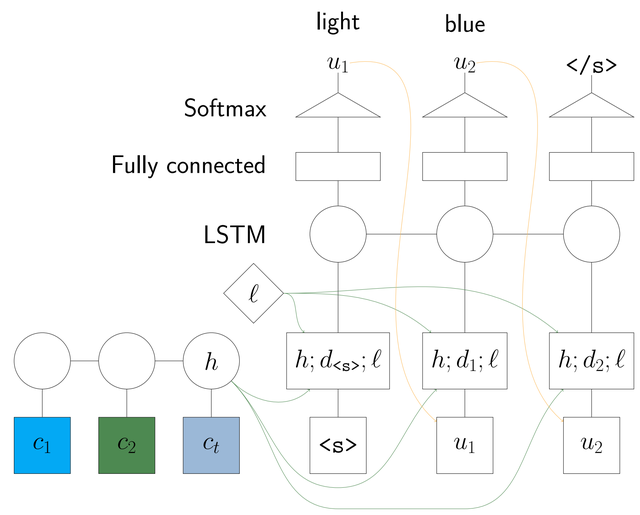
cocolab.stanford.edu/datasets/colors.html

Code is also publicly available:

github.com/futurulus/colors-in-context

Architecture

LSTM sequence-to-sequence model, concatenating previous output embedding with context representation and (if bilingual) 1-D embedding of “language bit” at each time step in decoding:



Metric

Pragmatic informativeness: how well would an ideal Bayesian listener do at interpreting human utterances, using the model to simulate a human speaker?

$$t \stackrel{?}{=} \arg\max_i S(u|c_i, \ell, C)$$

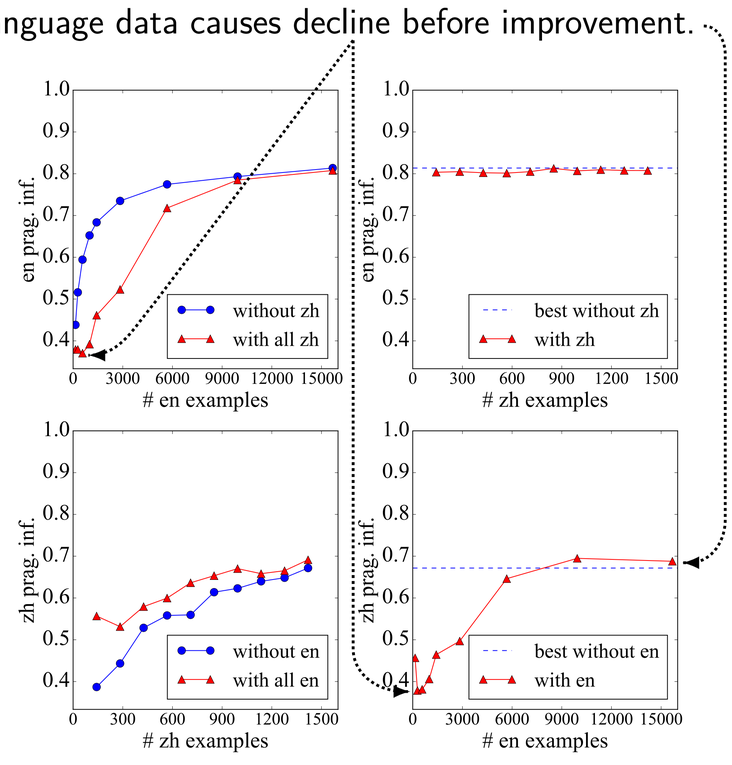
Results

Asymmetric benefits: English data helps dramatically on Chinese, Chinese data doesn't help on English.

test	train	dev acc	test acc
en	en	**80.51**	**83.06**
	en+zh	79.73	81.43
zh	zh	67.16	67.75
	en+zh	**71.81**	**72.89**

Effect of training data size

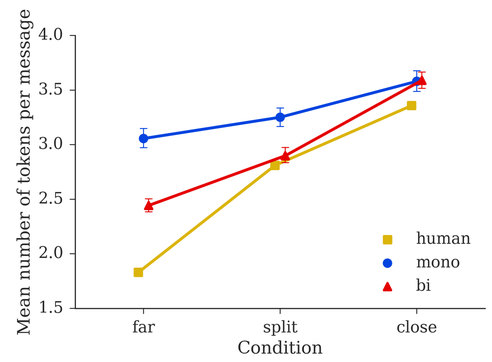
The English data is 10x larger than the Chinese data. Size difference is probably important: adding other-language data causes decline before improvement.



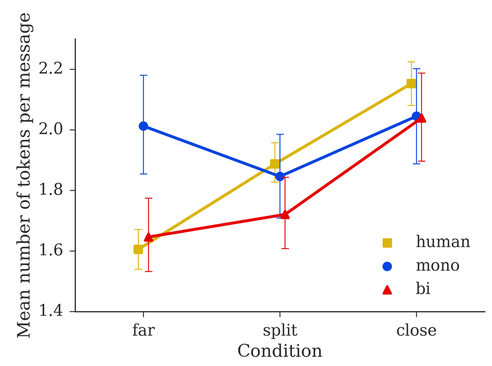
Human vs. models

Bilingual model is closer than monolingual to humans in corpus statistics of sampled utterances.

en



zh



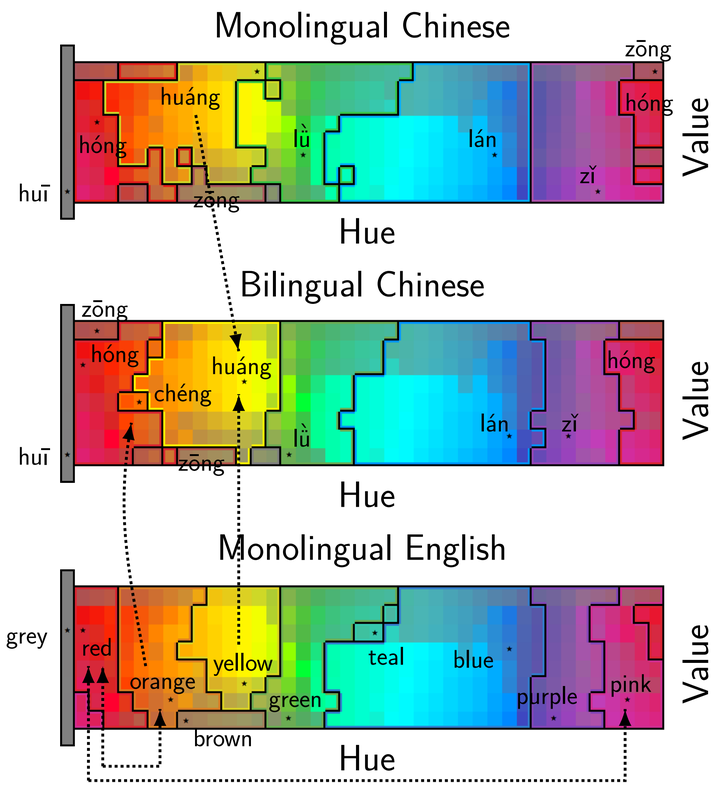
Bilingual lexicon induction

Using vector analogy and cosine nearest neighbor to identify most likely translations:

zh	→	en	en	→	zh
绿色	‘green’	**green**	green	绿	**‘green’**
紫色	‘purple’	**purple**	blue	蓝	**‘blue’**
蓝色	‘blue’	purple	purple	蓝	‘blue’
灰色	‘grey’	**grey**	bright	鲜艳	**‘bright’**
亮	‘bright’	**bright**	pink	粉色	**‘pink’**
灰	‘grey’	-er	grey	灰	**‘grey’**
蓝	‘blue’	*teal*	dark	暗	**‘dark’**
绿	‘green’	**green**	gray	灰	**‘grey’**
紫	‘purple’	**purple**	yellow	黄色	**‘yellow’**
草	‘grass’	*green*	light	最	‘most’

Color term semantics

We gave the World Color Survey task (Berlin & Kay, 1969) to the models, plotting the most likely utterance for each color averaged over random contexts.



Acknowledgments

We thank the Stanford Data Science Initiative & NSF for support. Poster template by Nathaniel Johnston.