eti

May 7, 2018

outline

team

Kuan Yu	kuanyu@uni-potsdam.de
Maya Angelova	maya.angelova@protonmail.com
Philipp Schoneville	schoneville@uni-potsdam.de

wikipedia datasets[1]

- aligned normal vs simple wikipedia content

version 1.0

- with train, tune, test splits

version 2.0

- no splits
- updated
- more data

[1] http://www.cs.pomona.edu/~dkauchak/simplification/

version 2.0 document-aligned

- 32 documents empty in normal or simple

simple

	min	max	mean
sent	1	916	8.46
word	0	10486	153.50
char	1	53497	806.37

normal

	min	max	mean
sent	1	1938	64.52
word	0	32026	1638.02
char	1	185194	8995.47

version 2.0 sentence-aligned

- 167,689 aligned sentence pairs

simple

	min	max	mean
word	1	192	22.86
char	17	2613	122.08

normal

	min	max	mean
word	1	236	25.55
char	17	2404	139.55

version 2.0 sentence-aligned

	normal	simple
`#word`	4284135	3832824
`#word-type`	162491	147078
`#non-hapax`	78494	71029
`#char-type`	133	134

- unique char `[` in sentence 152,563 did not unnormalized to `-LRB-` due to tokenization error

character-level modelling: byte-net[2]

- more robust
  - no `UNKNOWN`
  - no special treatment for numbers
  - may learn morphology
- no tokenizer required
- easier applicable to other languages

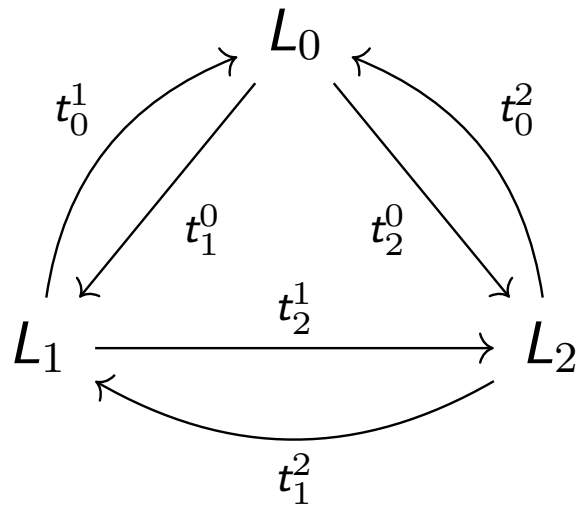
[2] `https://arxiv.org/abs/1610.10099`

attention: transformer-net[3]

- vs recurrent
  - faster training
  - no information bottleneck
- vs convolution
  - no limited input field

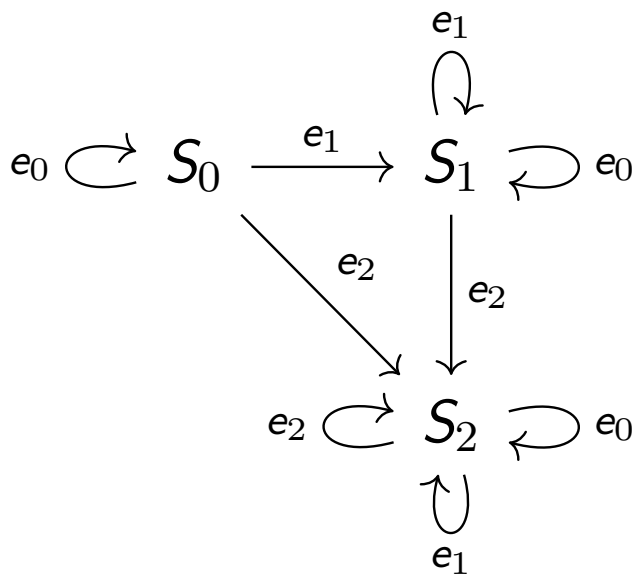
[3] https://arxiv.org/abs/1706.03762

concept art: the category of languages L



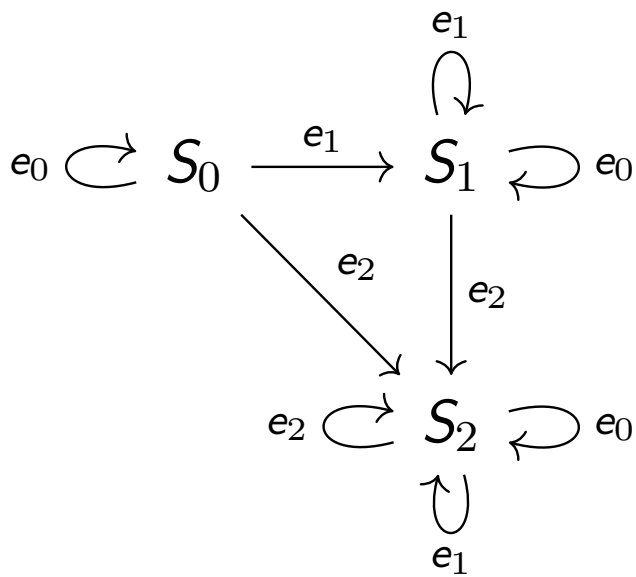
- $i, j, k \in \mathbb{N}$
- an object $L_i \in L$ is a language
- an arrow t_j^i translates $L_i \to L_j$, preserving discourse
- the composition $t_k^j t_j^i = t_k^i$ is an indirect translation

concept art: the category of a language L_i



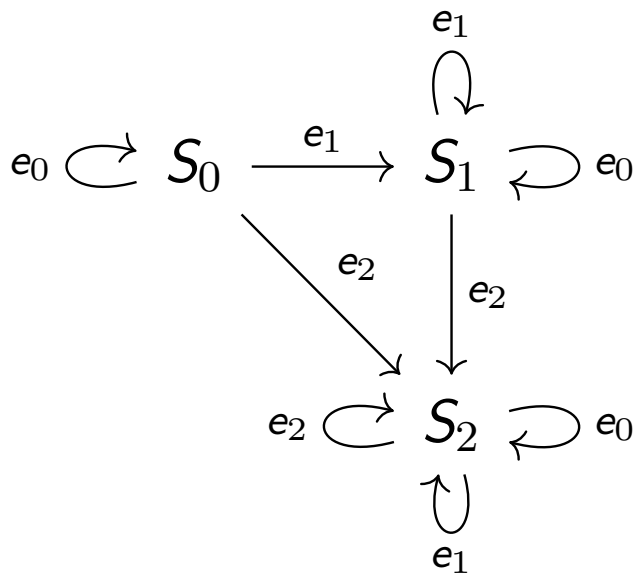
- $j, k, m, n \in \mathbb{N}$
- S_0 is the set of all possible worlds
- an object $S_j \in S = \mathbb{P}S_0$ is an information state
- an arrow $e_m : S_j \to S_k$ is an expression (proposition or utterance) which alters the information state
- the composition $e_n e_m = e_{mn}$, a concatenation of expressions, conducts discourse

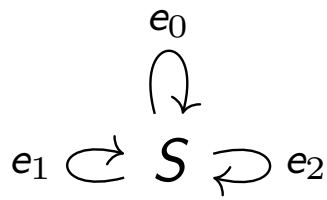
concept art: the category of a language L_i



- the empty expression e_0 adds no information to the discourse
- the intensional interpretation of e_1 is S_1, and e_2 S_2
- logically, e_2 implies e_1
- a gibberish leads any information state to the empty state, in which case the discourse must backtrack in order to proceed
- every state has one outgoing arrow corresponding to each expression (totality)

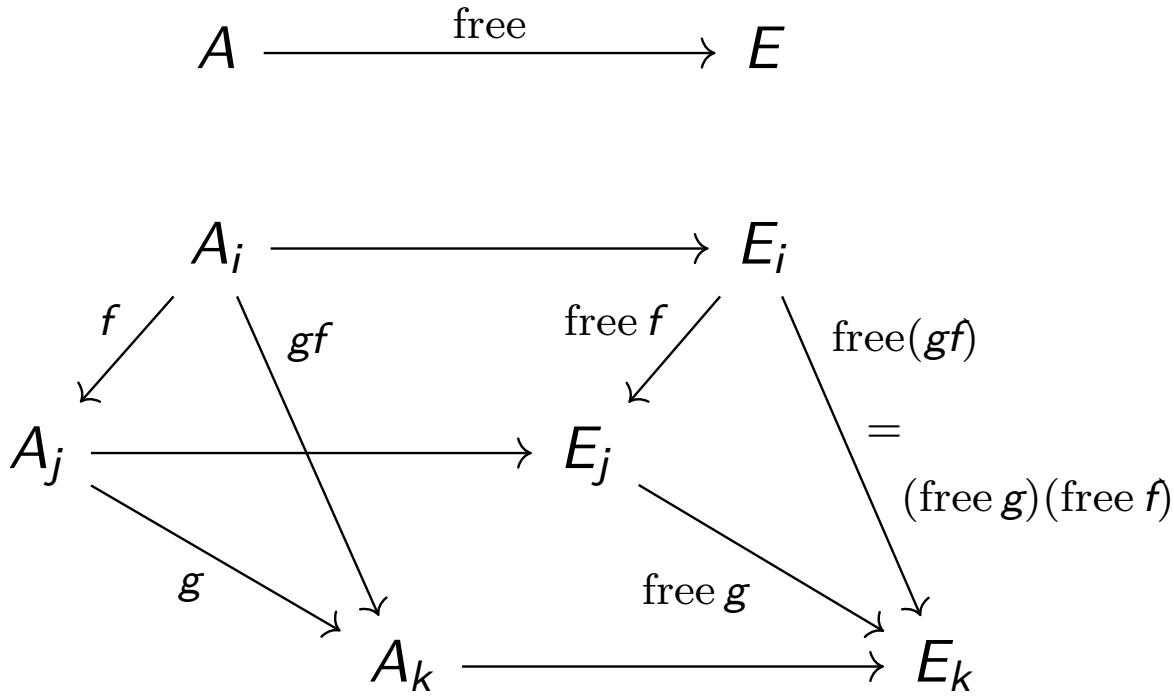
concept art: the monoid of a language E_i





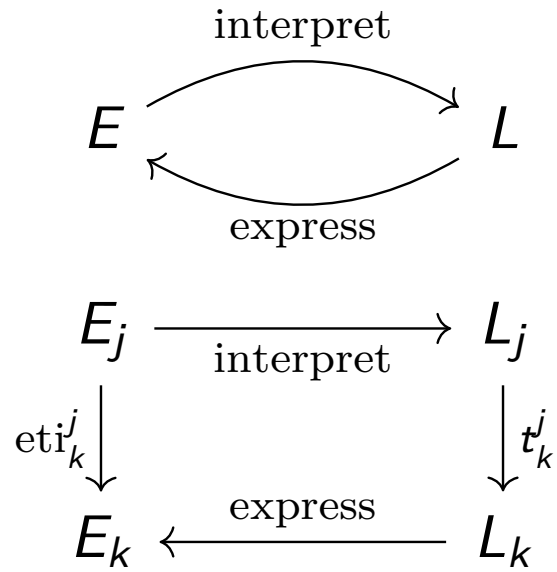
- L_i is a category with totality (left), which is a monoid (right)
- its composition is the concatenation of expressions
- its identity is the empty expression
- it is the free monoid generated by the alphabet A_i of L_i
- namely the free monoid $E_i = A_i^*$

concept art: the free functor $A \to E$



category	A	$E = \text{free}\, A$
objects	charsets	monoids
arrows	functions	monoid homomorphisms

concept art: eti



- the encoder functor $\mathrm{interpret}$ empirically identifies expressions in E_i as paths in L_i
- the decoder functor $\mathrm{express}$ reconstructs expressions from paths of discourse in L_i
- our model, the monoid homomorphism $\mathrm{eti}^j_k : E_j \to E_k$ is given by the composition $\mathrm{express}\, t^j_j \,\mathrm{interpret}$

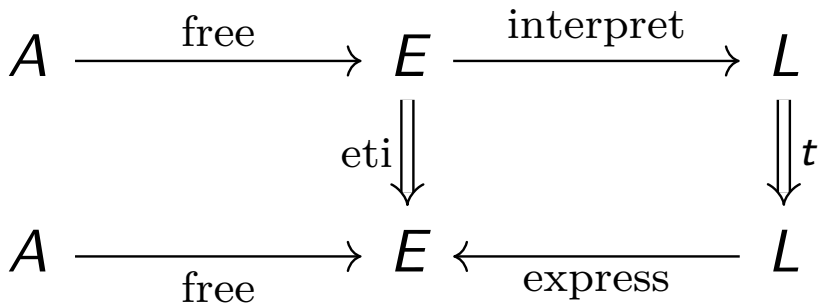
preparation

- unnormalized `-LRB- -RRB- -RCB- -LCB-`
- investigate and fix encoding error `\x92`
- train, tune, test splits

baselines

- run byte-net and transformer-net
- evaluate BLEU scores
  - bad measure but commonly used
  - consider ideas from other teams

modelling



exploration: CRF decoding

- instead of beam search
- use CRF log likelihood as loss
- how to fix sequence lengths?
  - consider the approach in conv-seq2seq[4]

[4] https://arxiv.org/abs/1705.03122

exploration: generative pretraining

- pretrain the encoder on the normal data, or
- pretrain the decoder on the simple data