Analysis and Visualization

Philipp Koehn

10 November 2020

analytical evaluation

Error Analysis

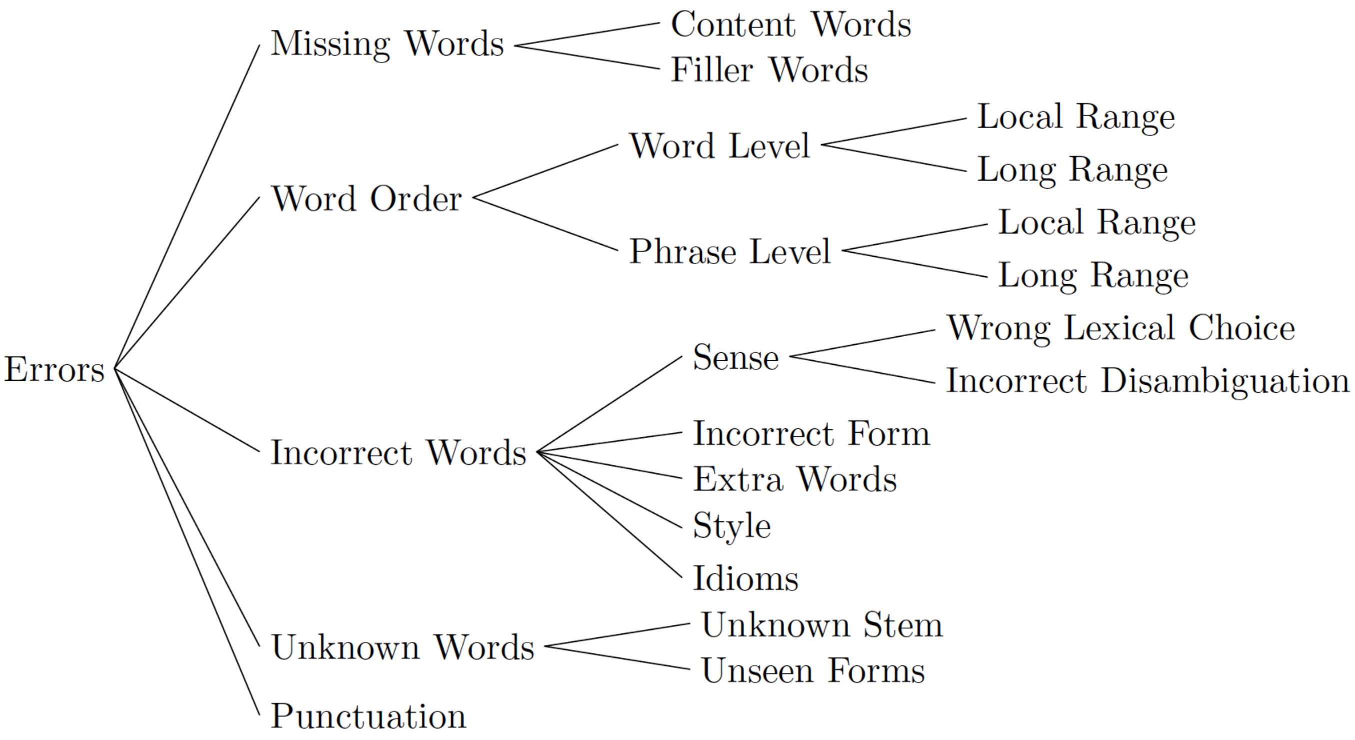
- Manually inspect output of machine translation system
- Identify errors and categorize them
- Specific problems of neural machine translation
  - dropped input / added output
  - gibberish (*the the the the*)
  - hallucinated output

Hallucinated Output

- Examples of extreme translation failures
  - Low resource example
    Republican strategy to counter the re-election of Obama
    Un órgano de coordinación para el anuncio de libre determinación
  - Out of domain example
    Schaue um dich herum.
    EMEA / MB / 049 / 01-EN-Final Work progamme for 2002

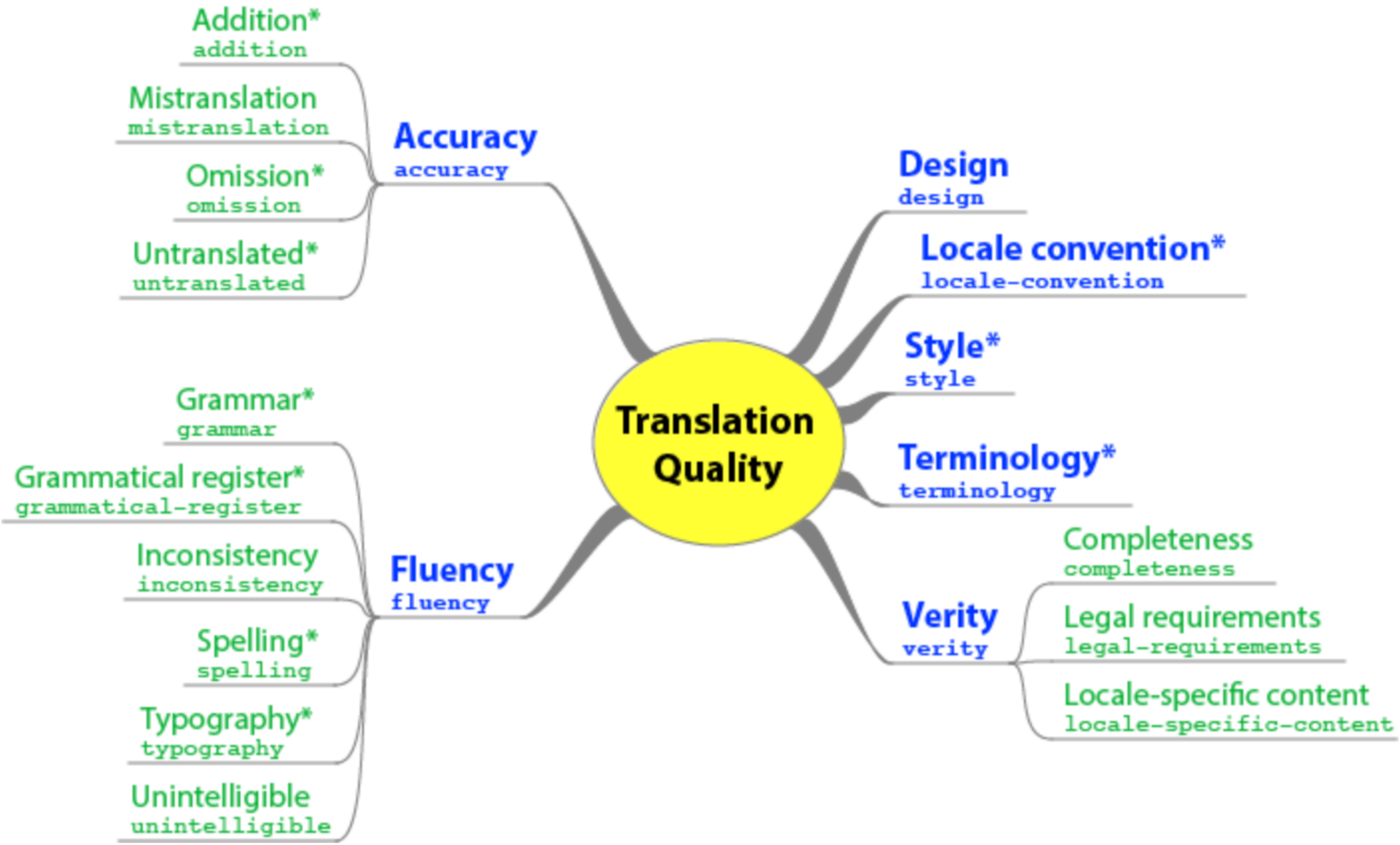
- Neural MT goes off track
  - turns into generative language model
  - ignores input context

Linguistic Categories



“Error Analysis of Statistical Machine Translation Output” (Vilar et al., LREC 2006)

MQM


Addition*
addition
Mistranslation
mistranslation
Omission*
omission
Untranslated*
untranslated
Accuracy
accuracy
Grammar*
grammar
Grammatical register*
grammatical-register
Inconsistency
inconsistency
Spelling*
spelling
Typography*
typography
Unintelligible
unintelligible
Fluency
fluency
Translation Quality
Design
design
Locale convention*
locale-convention
Style*
style
Terminology*
terminology
Verity
verity
Completeness
completeness
Legal requirements
legal-requirements
Locale-specific content
locale-specific-content

Bentivogli et al. (EMNLP 2016)

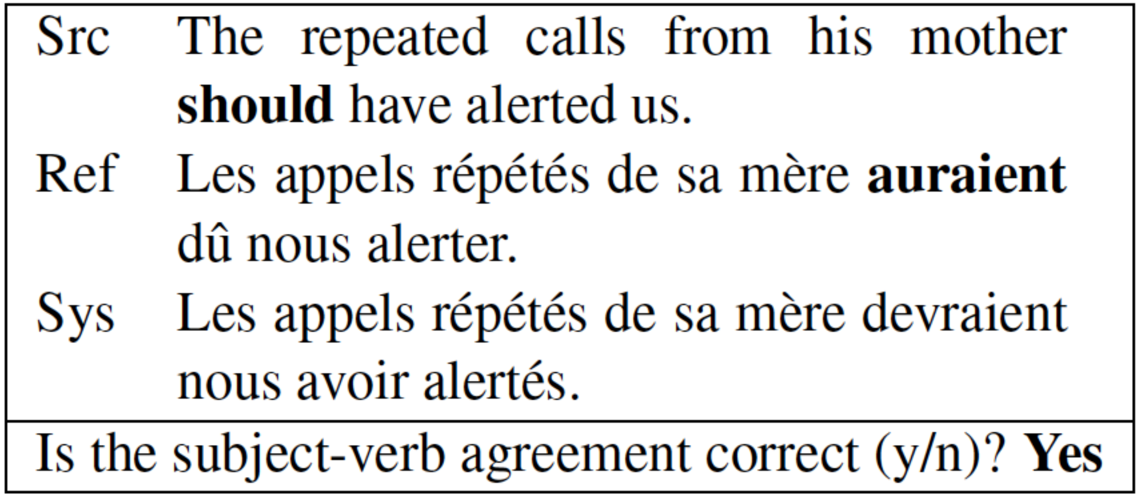
- Manually corrected machine translation
- Breakdown of word edits
  - by part-of-speech tag
  - multi-word construction, e.g., `AUX:V` constructions such as *can eat*
- Systems
  - NMT: neural machine translation
  - PBSY: phrase-based statistical
  - HPB: ierarchical phrase-based statistical
  - SPB: syntax-based statistical

Class	NMT-vs-PBSY	NMT	PBSY	HPB	SPB
V	-70%	35	116	133	155
PRO	-57%	22	51	53	62
PTKZU	-54%	6	13	4	11
ADV	-50%	14	28	44	36
N	-47%	37	70	99	56
KON	-33%	6	9	8	12
PREP	-18%	18	22	27	28
PTKNEG	-17%	10	12	10	7
ART	-4%	26	27	38	35
aux:V	-87%	3	23	17	18
neb:V	-83%	2	12	7	19
objc:V	-79%	3	14	21	24
subj:PRO	-70%	12	40	34	46
root:V	-68%	6	19	28	27
adv:ADV	-67%	8	24	33	28
obja:N	-65%	6	17	28	12
cj:V	-59%	7	17	21	22
part:PTKZU	-54%	6	13	4	11
obja:PRO	-38%	5	8	14	7
mroot:V	-36%	7	11	26	20
pn:N	-36%	16	25	33	19
subj:N	-33%	6	9	10	7
pp:PREP	-30%	14	20	19	23
adv:PTKNEG	-17%	10	12	10	7
det:ART	-4%	26	27	38	34
all	-48%	222	429	493	488

targeted test sets

Challenge Set

- Create challenging sentence pairs with specific problems



Src	The repeated calls from his mother **should** have alerted us.
Ref	Les appels répétés de sa mère **auraient** dû nous alerter.
Sys	Les appels répétés de sa mère devraient nous avoir alertés.
Is the subject-verb agreement correct (y/n)? **Yes**	

- “A Challenge Set Approach to Evaluating Machine Translation” (Isabelle et al., EMNLP 2017)

Challenge Set: Results

Category	Subcategory	#	PBMT-1	NMT	Google NMT
Morpho-syntactic	Agreement across distractors	3	0%	100%	100%
	through control verbs	4	25%	25%	25%
	with coordinated target	3	0%	100%	100%
	with coordinated source	12	17%	92%	75%
	of past participles	4	25%	75%	75%
	Subjunctive mood	3	33%	33%	67%
Lexico-syntactic	Argument switch	3	0%	0%	0%
	Double-object verbs	3	33%	67%	100%
	Fail-to	3	67%	100%	67%
	Manner-of-movement verbs	4	0%	0%	0%
	Overlapping subcat frames	5	60%	100%	100%
	NP-to-VP	3	33%	67%	67%
	Factitives	3	0%	33%	67%
	Noun compounds	9	67%	67%	78%
	Common idioms	6	50%	0%	33%
	Syntactically flexible idioms	2	0%	0%	0%
Syntactic	Yes-no question syntax	3	33%	100%	100%
	Tag questions	3	0%	0%	100%
	Stranded preps	6	0%	0%	100%
	Adv-triggered inversion	3	0%	0%	33%
	Middle voice	3	0%	0%	0%
	Fronted should	3	67%	33%	33%
	Clitic pronouns	5	40%	80%	60%
	Ordinal placement	3	100%	100%	100%
	Inalienable possession	6	50%	17%	83%
	Zero REL PRO	3	0%	33%	100%

Contrastive Translation Pairs

- Goal: find out how well certain translation problems are handled
- Examples
  - noun phrase agreement
  - subject-verb agreement
  - separable verb particle
  - polarity (negative/positive)
- Idea: forced decoding with contrastive translation pair
  - positive example: correct translation
  - negative example: translation with error
- Check if positive example gets better score

Contrastive Translation Pairs

- Noun phrase agreement
  - good: ... *these interesting proposals* ...
  - bad: ... *this interesting proposals* ...
- Subject-verb agreement
  - good: ... *the idea to extend voting rights was* ...
  - bad: ... *the idea to extend voting rights were* ...
- Separable verb prefix
  - good: ... *switch the light on* ...
  - bad: ... *switch the light by* ...

Sennrich (EACL 2017)

- Compares neural machine translation systems for English–German
- Varying word encoding
  - byte pair encoding (BPE)
  - character-based word embeddings (char)
- Results

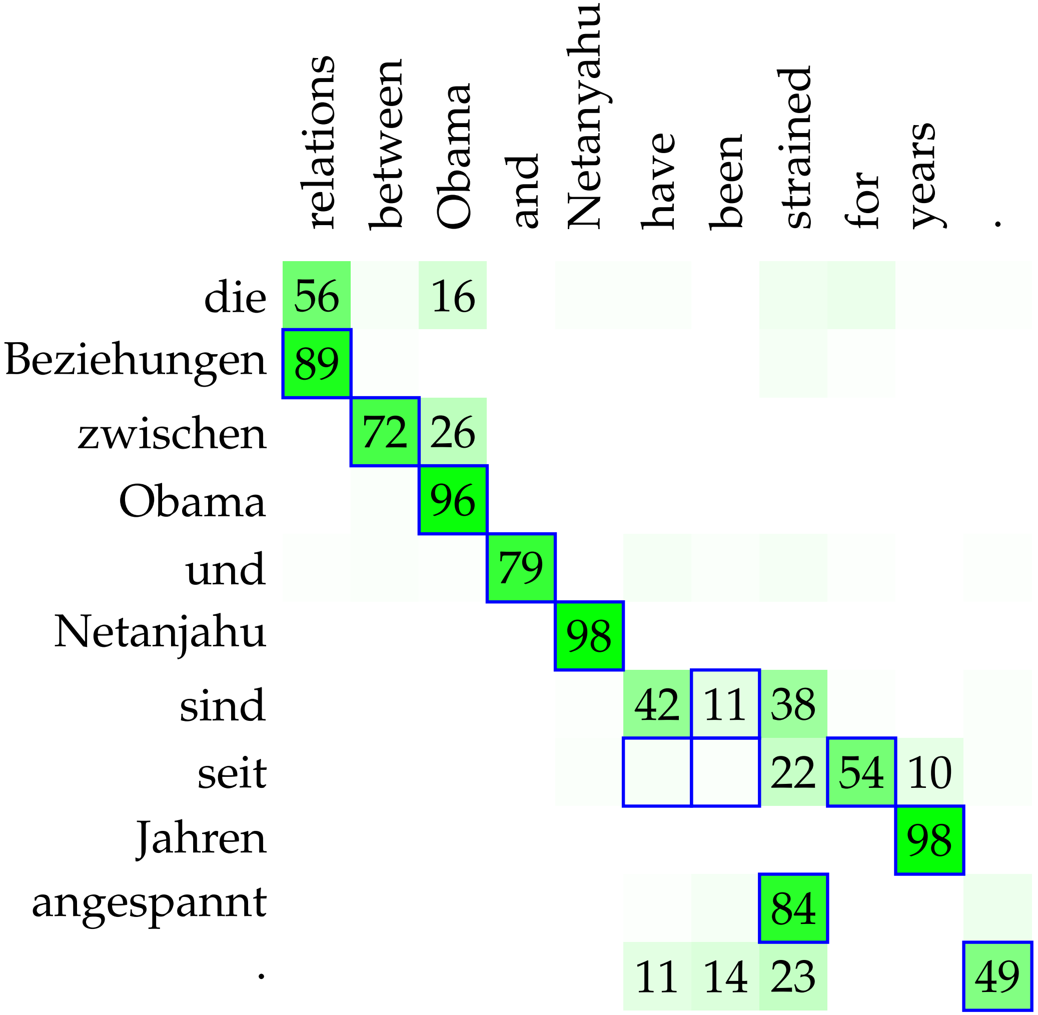
	agreement			polarity (negation)	
system	noun phrase	subject-verb	verb particle	insertion	deletion
BPE-to-BPE	95.6	93.4	91.1	97.9	91.5
BPE-to-char	93.9	91.2	88.0	98.5	88.4
char-to-char	93.9	91.5	86.7	98.5	89.3
human	99.4	99.8	99.8	99.9	98.5

Synthetic Languages

- Create artificial training examples to assess capability of systems
- Example: bracketing language
  - ({ })
  - ({ } { ( ) })
  - { ({ } ({ })) ({ }) }
- Check ability to make correct predictions based on nesting depth, length, etc.

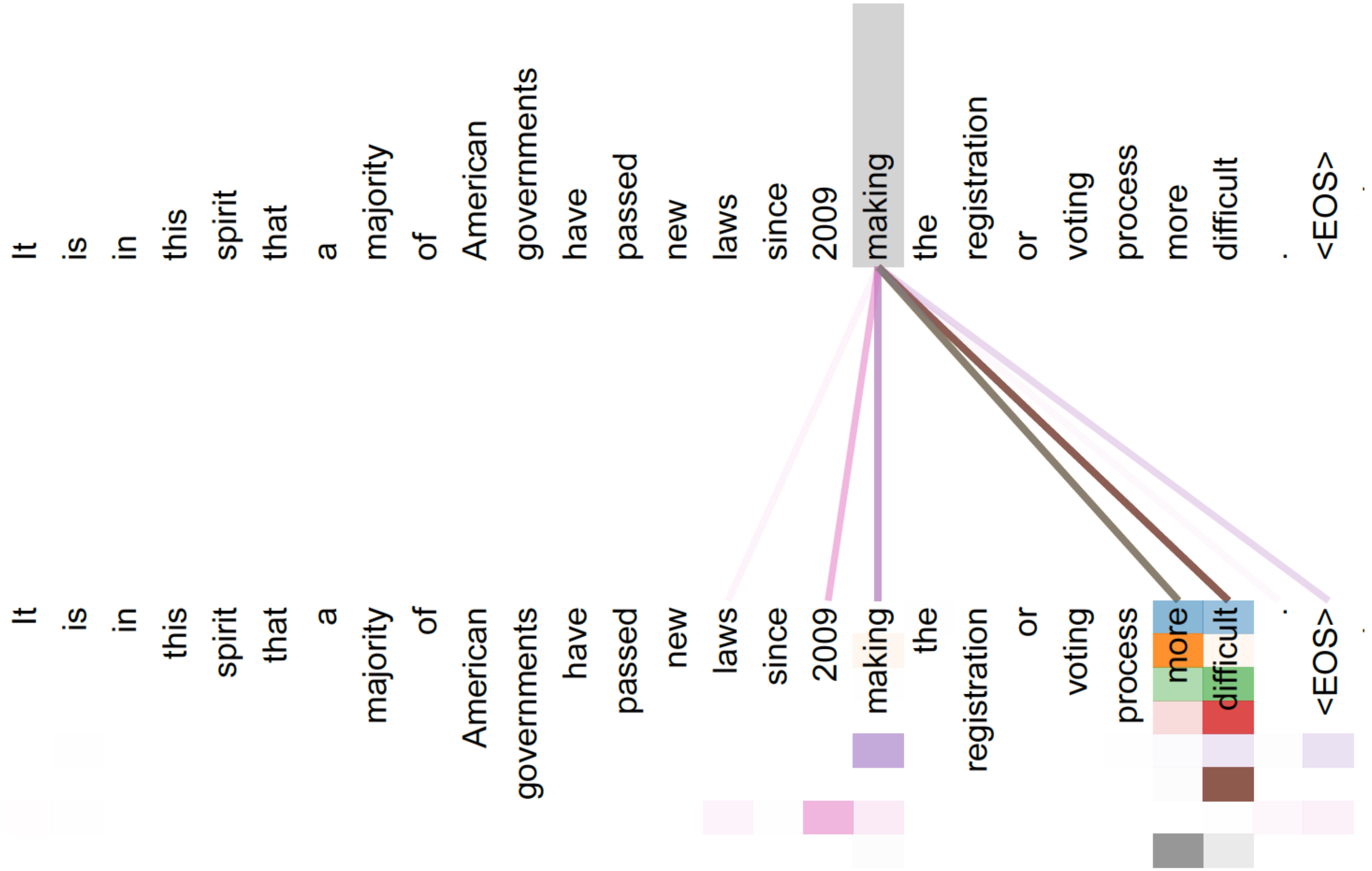
visualization

Word Alignment

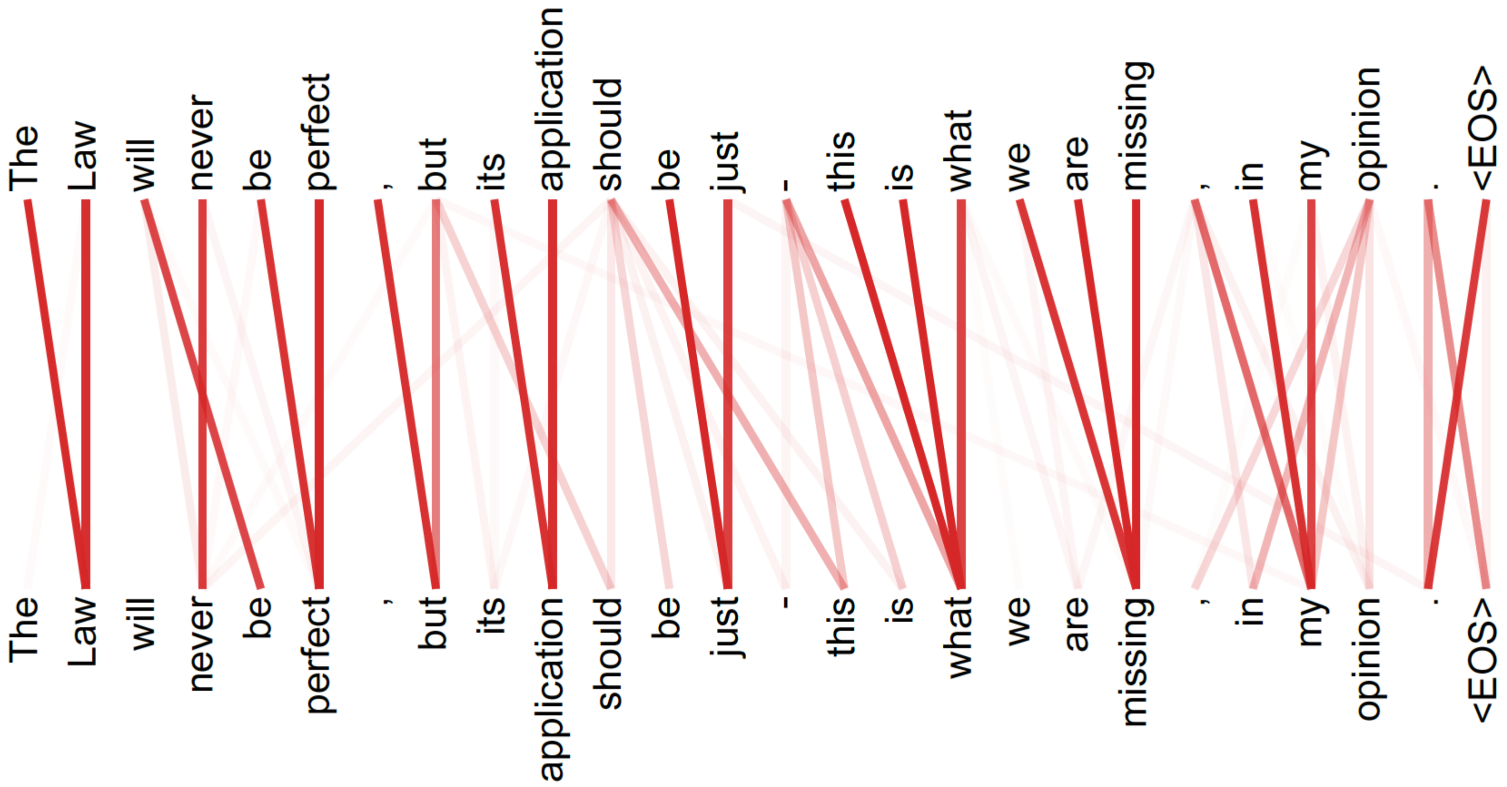


	relations	between	Obama	and	Netanyahu	have	been	strained	for	years	.
die	56		16								
Beziehungen	89										
zwischen		72	26								
Obama			96								
und				79							
Netanjahu					98						
sind						42	11	38			
seit								22	54	10	
Jahren										98	
angespannt								84			
.						11	14	23			49

Multi-Head Attention

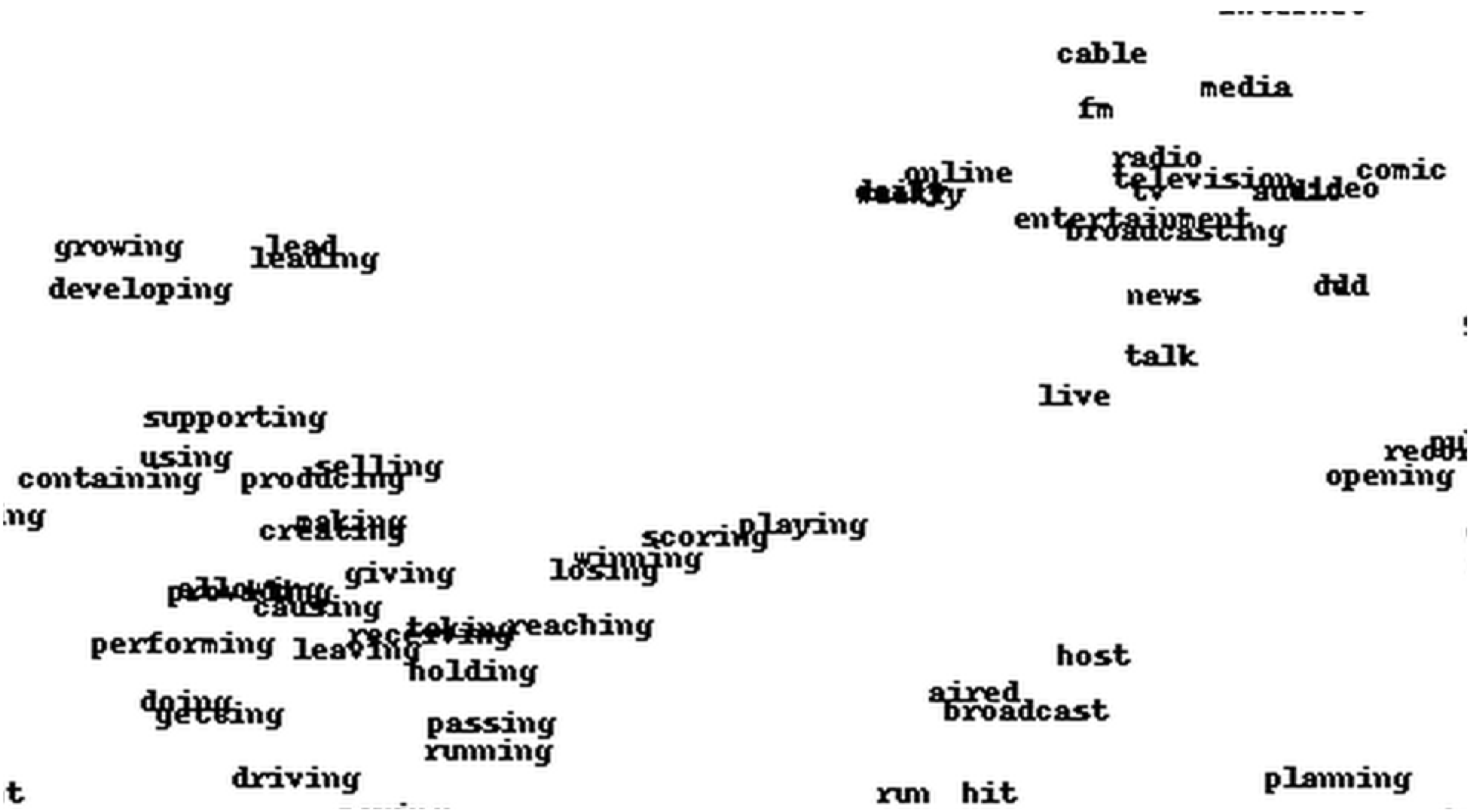


Multi-Head Attention

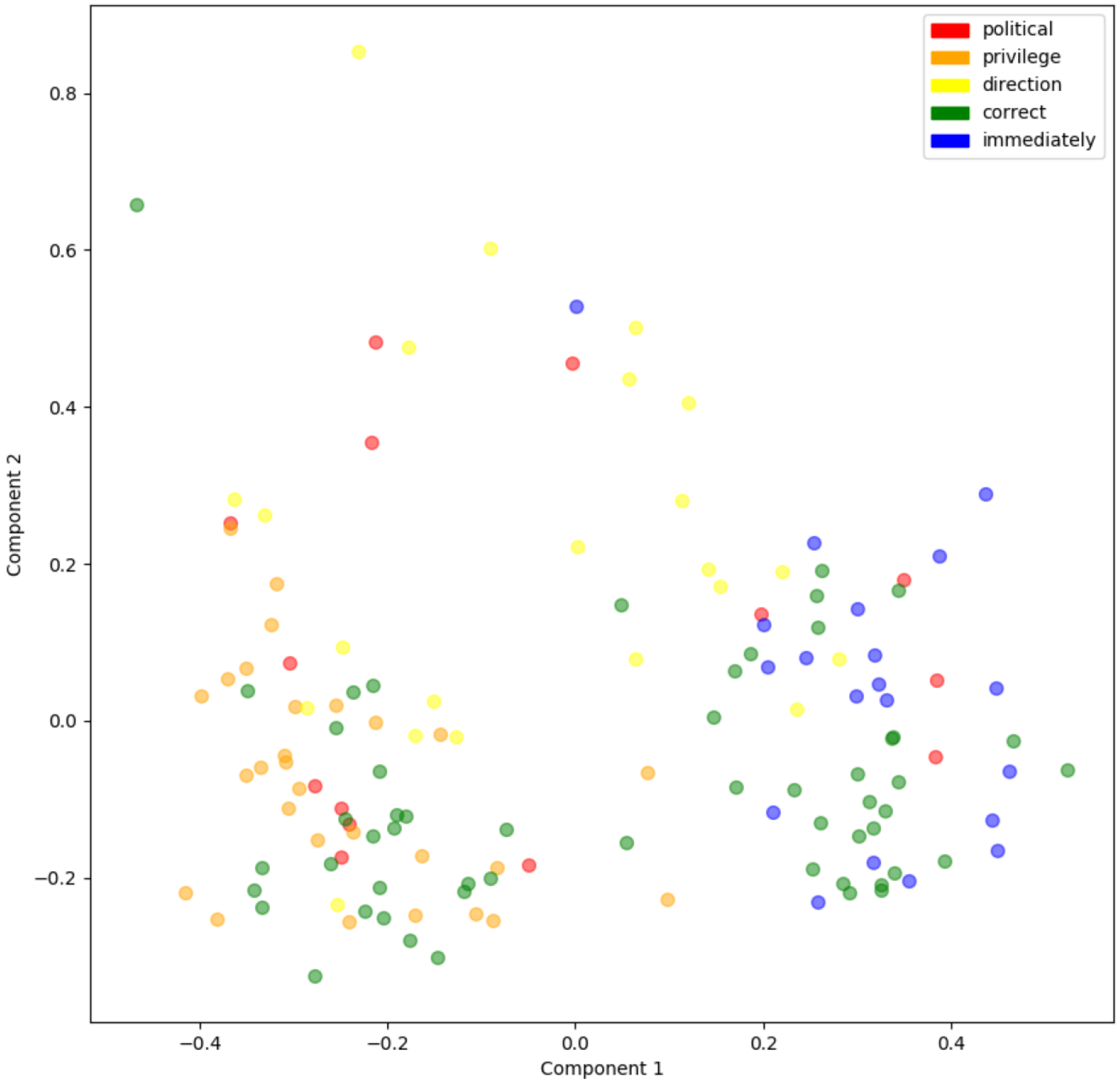


“Many of the attention heads exhibit behaviour that seems related to the structure of the sentence.“

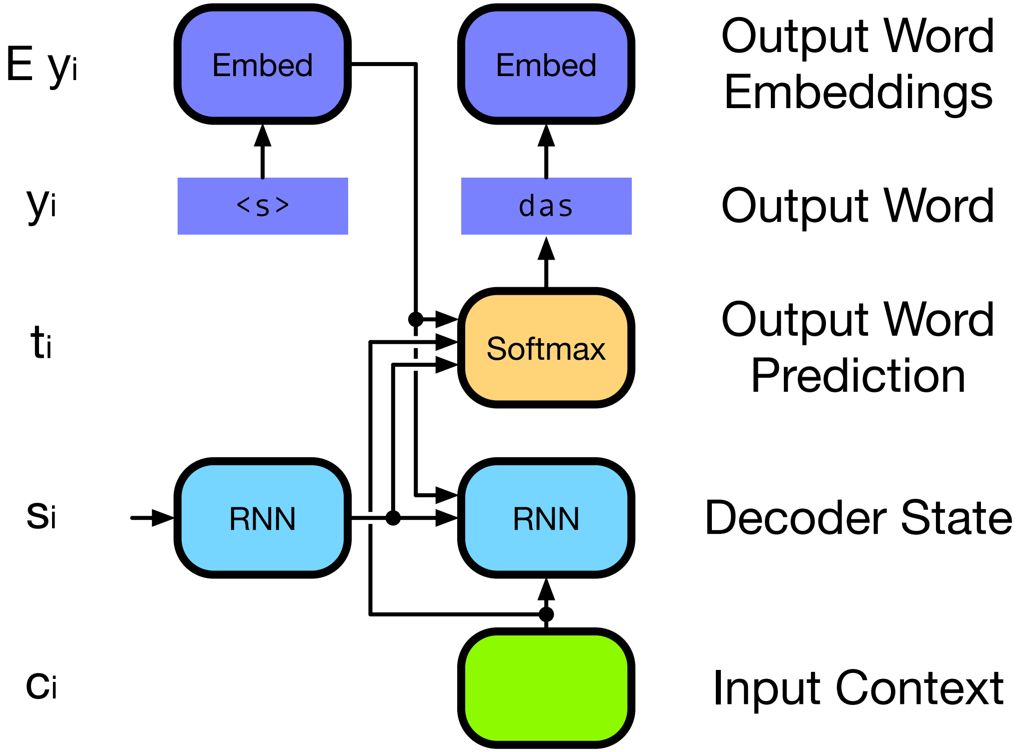
Word Embeddings



Word Sense Clusters



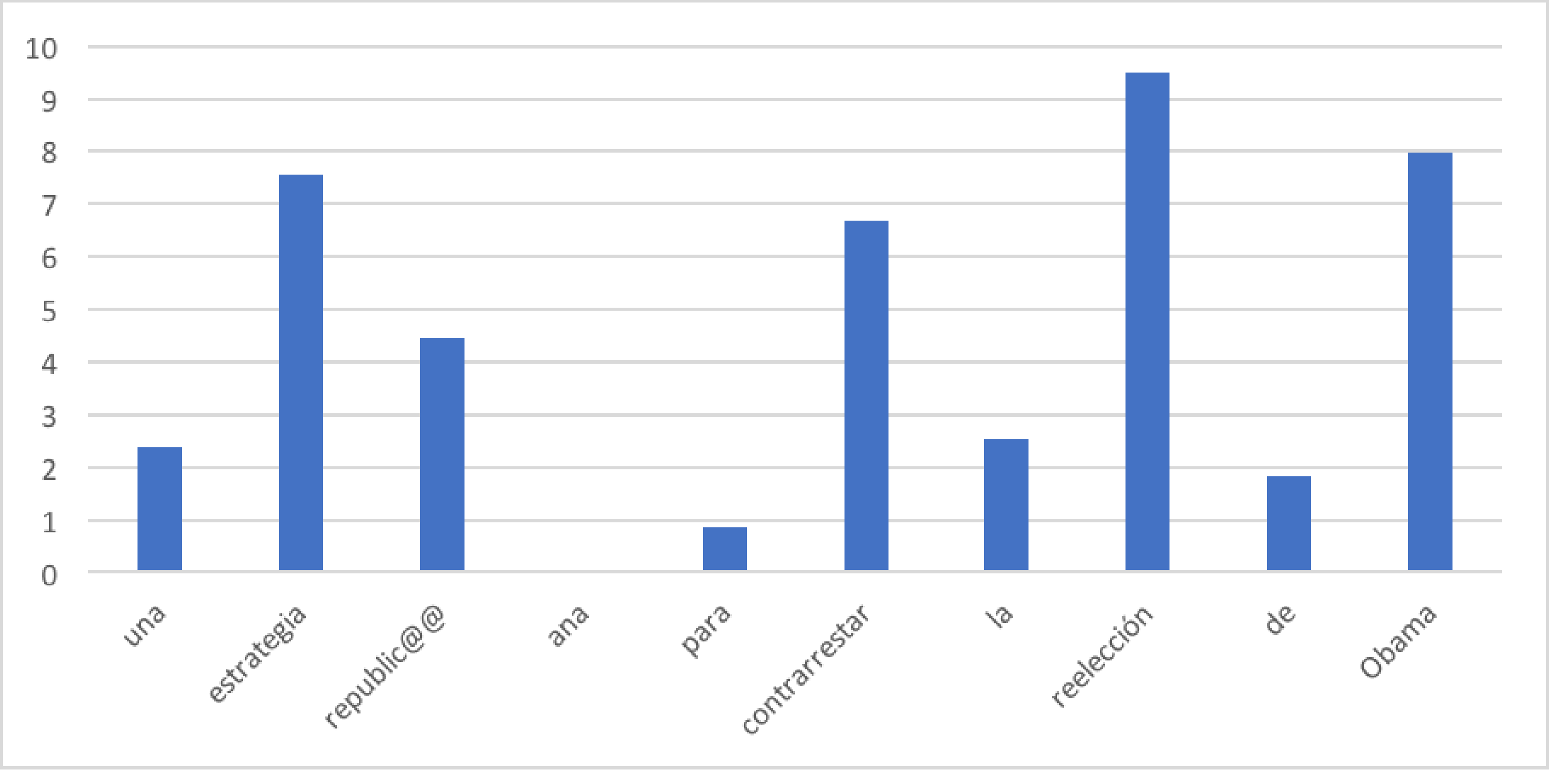
Input Context and Decoder State



- Word predictions are informed by previous output (decoder state) and input
- How much does each contribute?

Input Context vs. Decoder State

- Input: *Republican strategy to counter the re @-@ election of Obama*



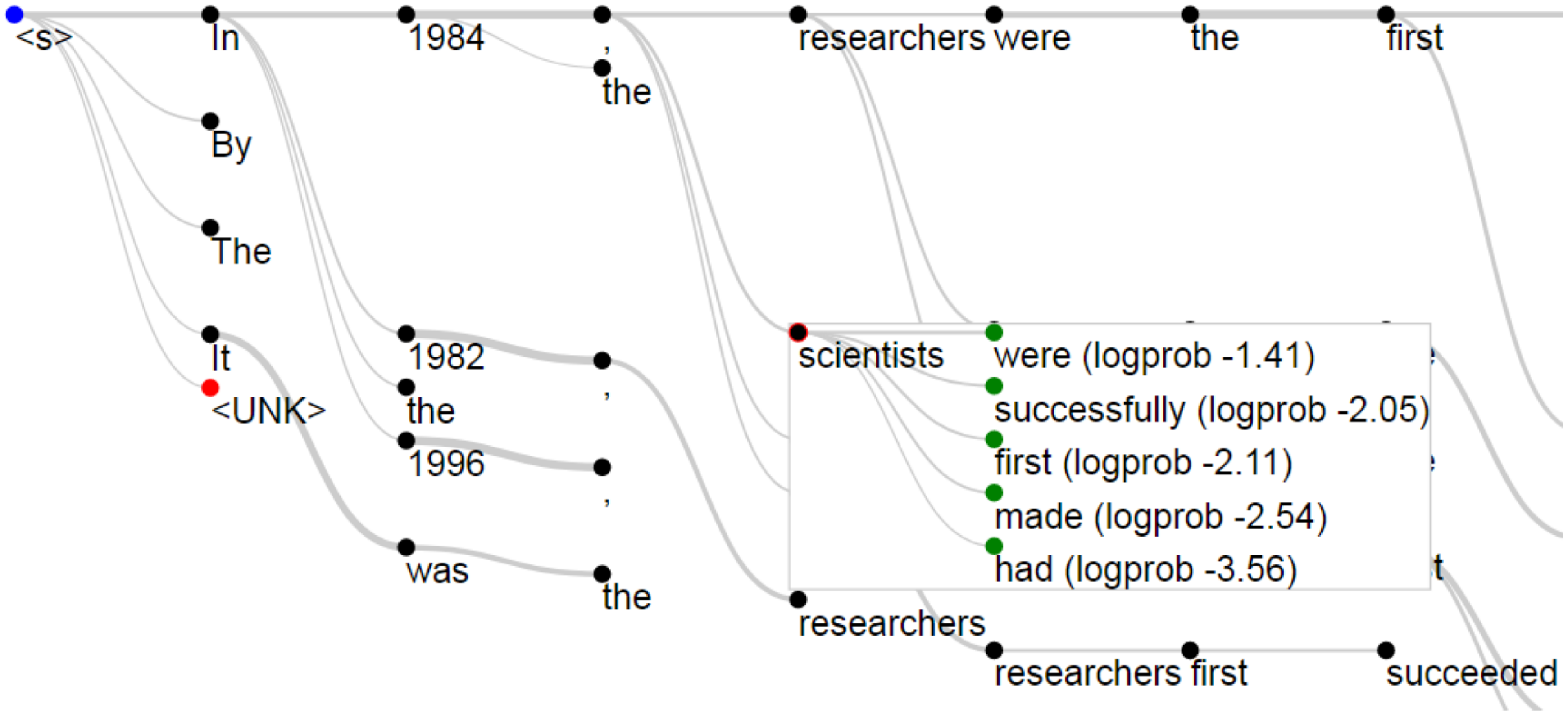
- KL divergence between decoder predictions with and w/o input context
- Input context matters more for content words

visualization tools

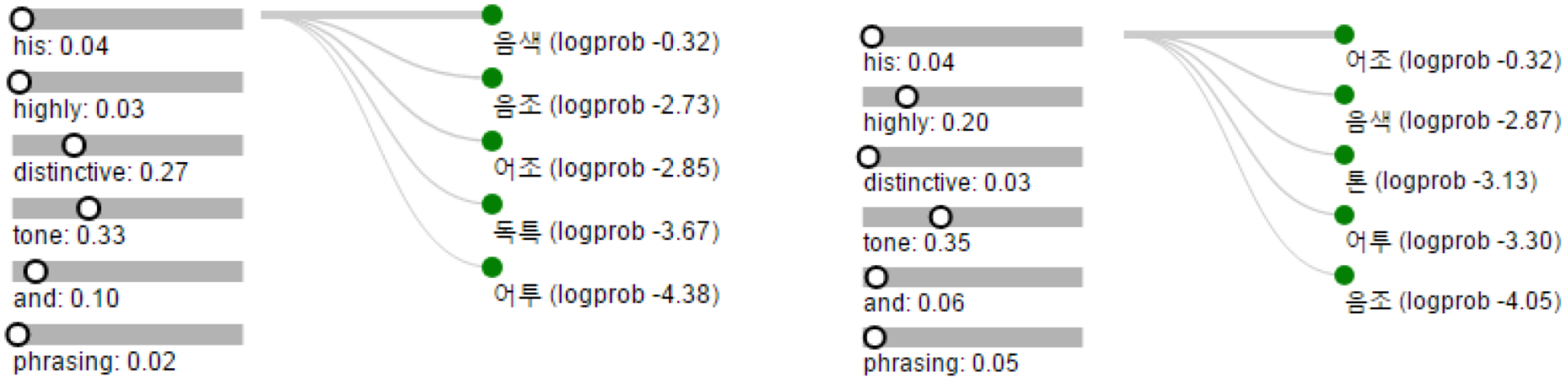
Interactive Exploration

- Tools for inspecting behavior of models and algorithms
- Helps to get insights
- Examples
  - "Interactive Visualization and Manipulation of Attention-based Neural Machine Translation" (Lee et al., EMNLP 2017)
  - "SEQ2SEQ-VIS : A Visual Debugging Tool for Sequence-to-Sequence Models" (Strobelt et al., 2018)

Search Graph

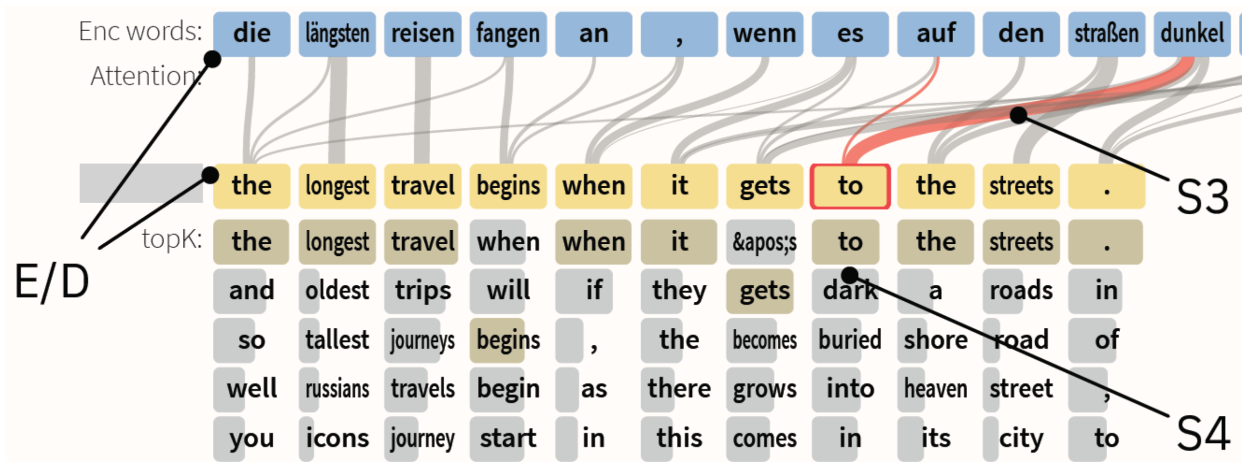


Manipulating Search



- Inspect attention weights
- Change attention weights → check change in word prediction

Predictions



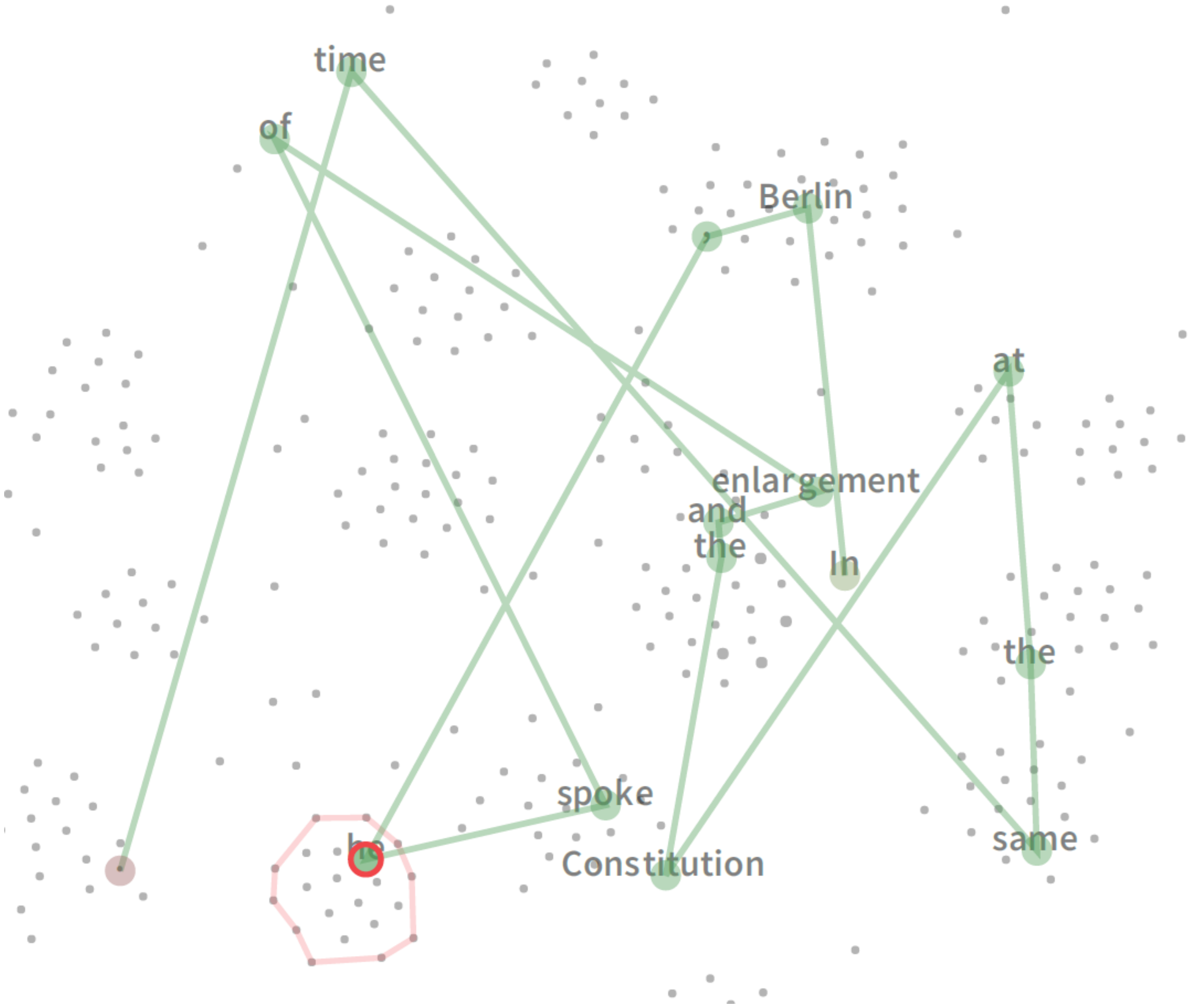
Enc words: die längsten reisen fangen an , wenn es auf den straßen dunkel wird .

the longest travel begins when it gets to the streets .

the	longest	travel	when	when	it	&apos;s	to	the	streets	.
and	oldest	trips	will	if	they	gets	dark	a	roads	in
so	tallest	journeys	begins	,	the	becomes	buried	shore	road	of
well	russians	travels	begin	as	there	grows	into	heaven	street	,
you	icons	journey	start	in	this	comes	in	its	city	to

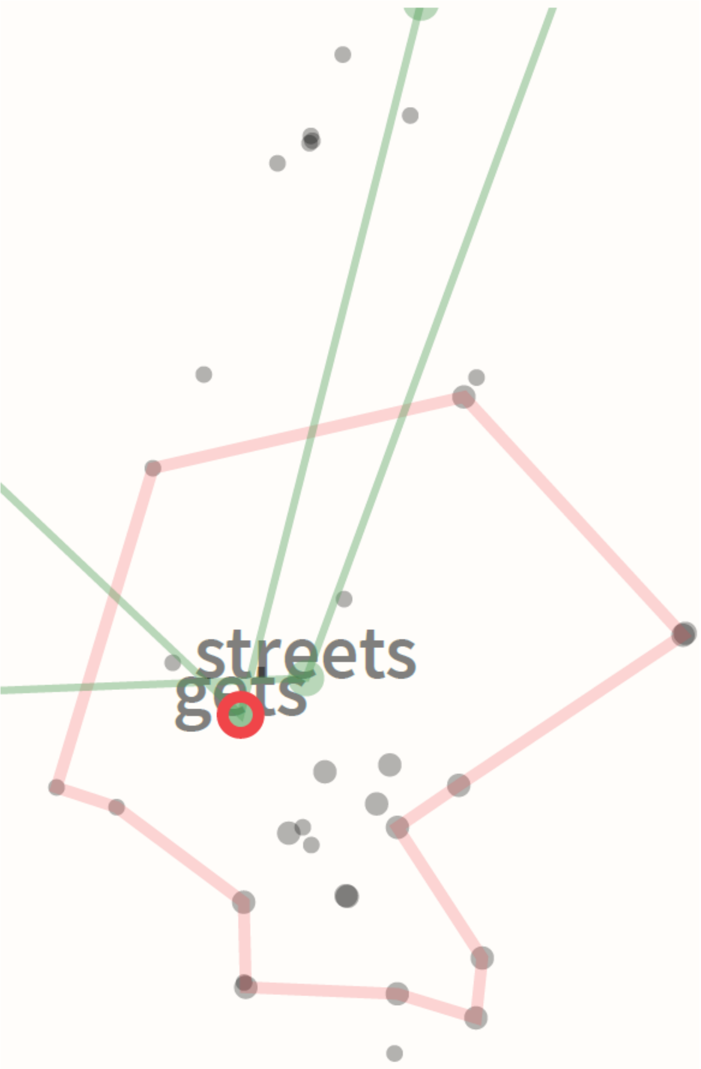
- E/D: encoder and decoder words
- S3: attention weights
- S4: top k predictions

Trajectory of Decoder States

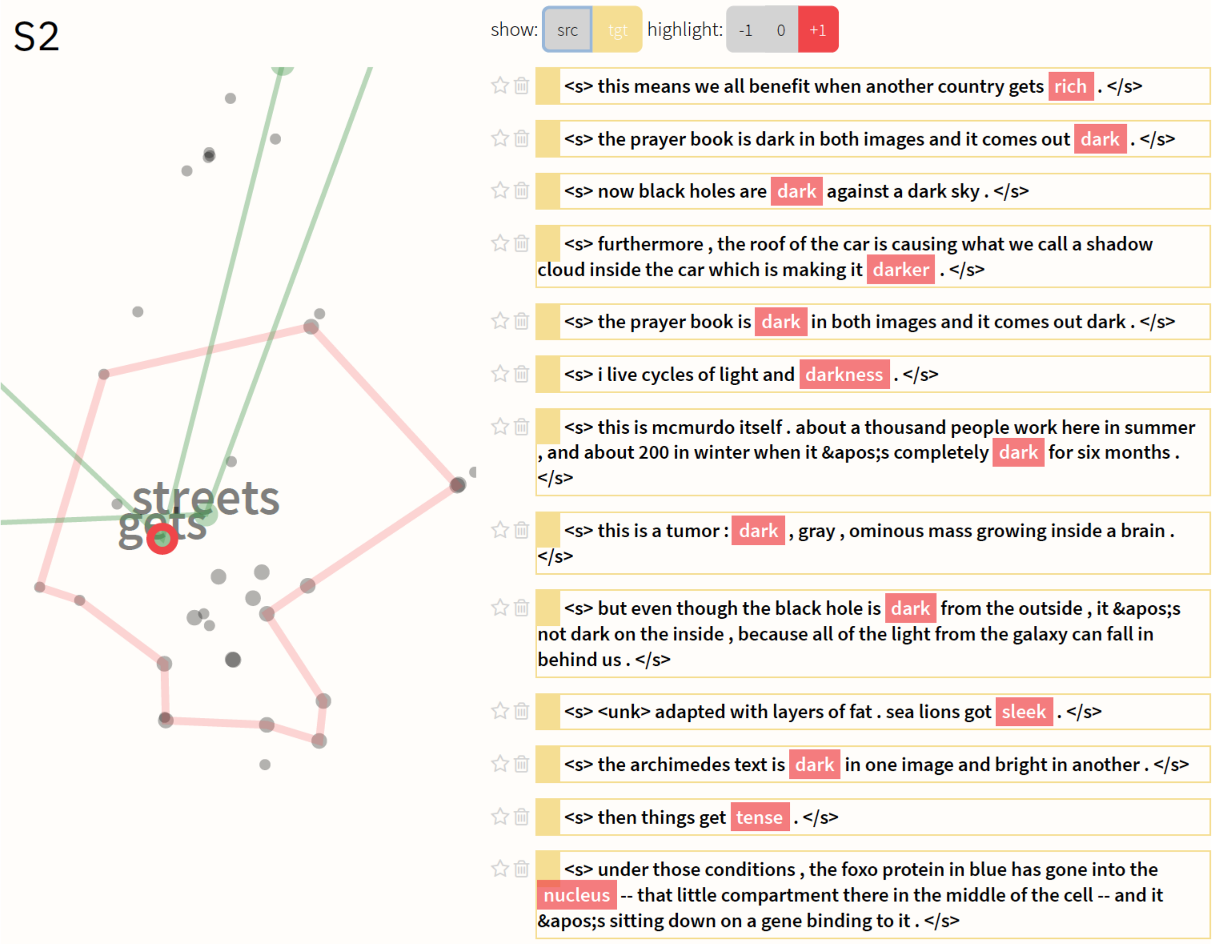


Decoder State Neighborhoods

- 2-D projections of decoder states
- Database of decoder states in training data
- Show neighborhood



Similar Decoder State



<s> this means we all benefit when another country gets rich . </s>

<s> the prayer book is dark in both images and it comes out dark . </s>

<s> now black holes are dark against a dark sky . </s>

<s> furthermore , the roof of the car is causing what we call a shadow cloud inside the car which is making it darker . </s>

<s> the prayer book is dark in both images and it comes out dark . </s>

<s> i live cycles of light and darkness . </s>

<s> this is mcmurdo itself . about a thousand people work here in summer , and about 200 in winter when it &apos;s completely dark for six months . </s>

<s> this is a tumor : dark , gray , ominous mass growing inside a brain . </s>

<s> but even though the black hole is dark from the outside , it &apos;s not dark on the inside , because all of the light from the galaxy can fall in behind us . </s>

<s> <unk> adapted with layers of fat . sea lions got sleek . </s>

<s> the archimedes text is dark in one image and bright in another . </s>

<s> then things get tense . </s>

<s> under those conditions , the foxo protein in blue has gone into the nucleus -- that little compartment there in the middle of the cell -- and it &apos;s sitting down on a gene binding to it . </s>

probing representations

What is in a Representation?

- What is contained in an intermediate representation?
  - word embedding
  - encoder state
  - decoder state

- More specific questions
  - does the model discover morphological properties?
  - does the model disambiguate words?

Classifier Approach

- Pose a hypothesis, e.g.,

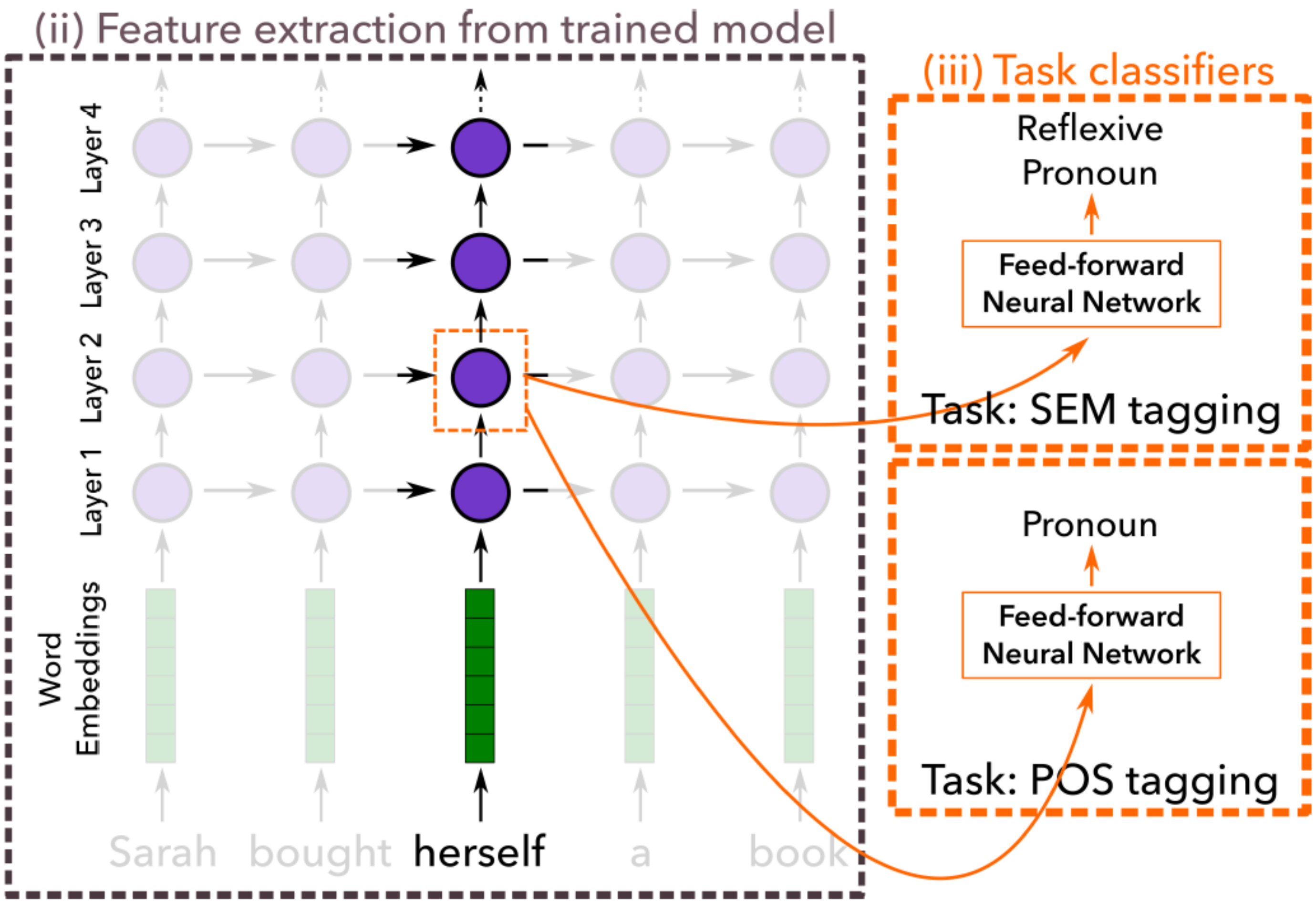
Encoder states discover part-of-speech.

- Formalize this as a classification problem
  - given: encoder state for word *dog*
  - label: singular noun (NN)

- Train on representations generated by running inference
  - translate sentences not seen during training
  - record their encoder states
  - look up their part of speech tags (running POS tagger or use labeled data)

  → training example (encoder state ; label)

- Test on new sentences



"Evaluating Layers of Representation in Neural Machine Translation on Part-of-Speech and Semantic Tagging Tasks" (Belinkov et al., ACL 2017)

Shi et al. (EMNLP 2016)

- LSTM sequence-to-sequence model without attention

- Different tasks
  - translate English into Russian, German
  - copy English
  - copy permuted English
  - parse English into linearized parse structure

- Predict
  - constituent phrase (NP, VP, etc.)
  - passive voice and tense

- Findings
  - much better quality when translating than majority class
  - same quality for copying as majority class

Belinkov et al. (EMNLP 2017)

- Attentional neural machine translation model

- Predict
  - part-of-speech tag
  - semantic tag
    - type of named entity
    - semantic relationships
    - discourse relationships

- Findings
  - compare prediction quality of different encoder layers
  - mostly better performance at deeper layers
  - little impact from target language

Belinkov et al. (ACL 2017)

- Attentional neural machine translation model with character-based word embeddings

- Predict for morphologically rich input languages
  - part-of-speech tag
  - morphological properties

- Findings
  - character-based representations much better for learning morphology
  - word-based models are sufficient for learning structure of common words
  - lower layers better at word structure, deeper layers better at word meaning
  - target language matters for what information is learned
  - neural decoder learns very little about word structure

relevance propagation

What Determined Output Decision?

- What part of the network had the biggest impact on final decision?
- For instance machine translation:
  - prediction of a specific output word
  - which of the input words mattered most?
  - which of the previous output words mattered most?

Layer-Wise Relevance Propagation

- Start with output prediction

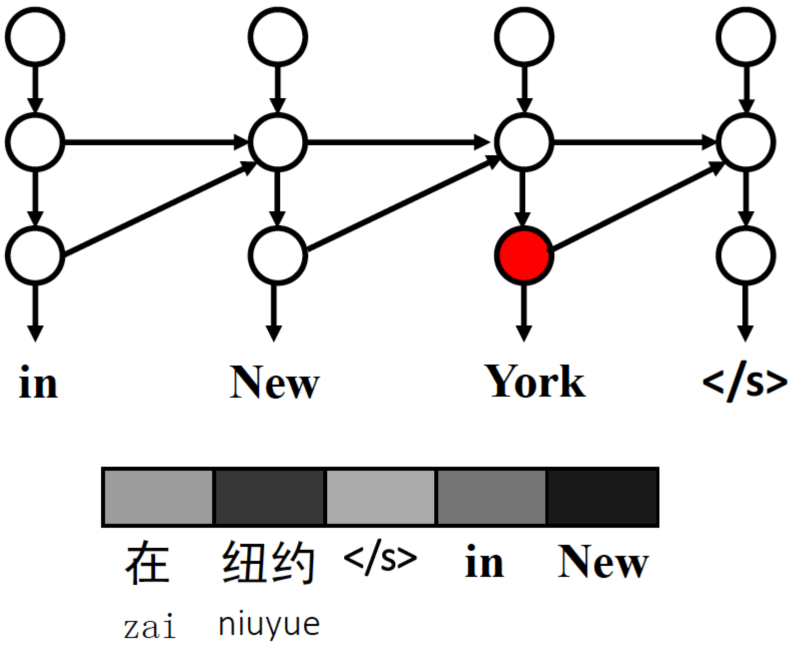
  i.e., high value for word in softmax

- Compute backwards what contributed to this high value

- First step
  - consider values of previous layer
  - consider weights from previous layer
  - assign relevance values for each node in previous layer
  - normalize so they add up to one

- Recurse until input layer is reached

Example: Chinese–English



"Visualizing and Understanding Neural Machine Translation"
(Ding, Liu, Luang and Sun, ACL 2017)

saliency

Saliency

- Intuition

if a decision changes a lot if a specific input value changes

$\Downarrow$

more relevant▮

change in the input value has no impact on decision

$\Downarrow$

not relevant▮

- Mathematically
  - relationship $p(y_0|x_0)$ between an input value x_0 and an output value y_0
  - assume this to be a linear relationship (which is approximately true locally)
  - compute slope by derivative

$$\text{saliency}(x, y) = \frac{\partial p(x|y)}{\partial x}$$

Example: Word Alignment

- Which input word had the most influence on an output word prediction?

⇒ Trace back to word embeddings

- Note
  - not interested in individual neurons
  - combine salience values in embedding vector

Saliency

	we	do	not	believe	that	we	should	cherry-pick	.
wir	■								
glauben		■		■					
nicht			■						
,					■				
daß					■				
wir						■			
nur								■	
rosinen								■	
herauspicken								■	
sollten							■		
.									■

Human Reference

	we	do	not	believe	that	we	should	cherry-pick	.
wir	■								
glauben				■					
nicht		■	■						
,									
daß									
wir					■	■			
nur							■	■	
rosinen								■	
herauspicken								■	■
sollten									
.									

Attention

	we	do	not	believe	that	we	should	cherry-pick	.
wir	■								
glauben				■					
nicht		■	■						
,									
daß					■				
wir						■			
nur									
rosinen								■	
herauspicken								■	
sollten							■		
.									■

Saliency

What Do Interpretability Measures Reveal?

- How are do we know if these methods are doing the right thing

what a model should be doing
$\neq$
what a model is doing

- Also: impact of input word $\neq$ word alignment
  - *bank* most responsible to produce German translation *Bank*
  - *credit* or *account* may be crucial for word sense disambiguation
  - other words may provide clues that word is a noun (not a verb)

Explainable AI

- Important question for users

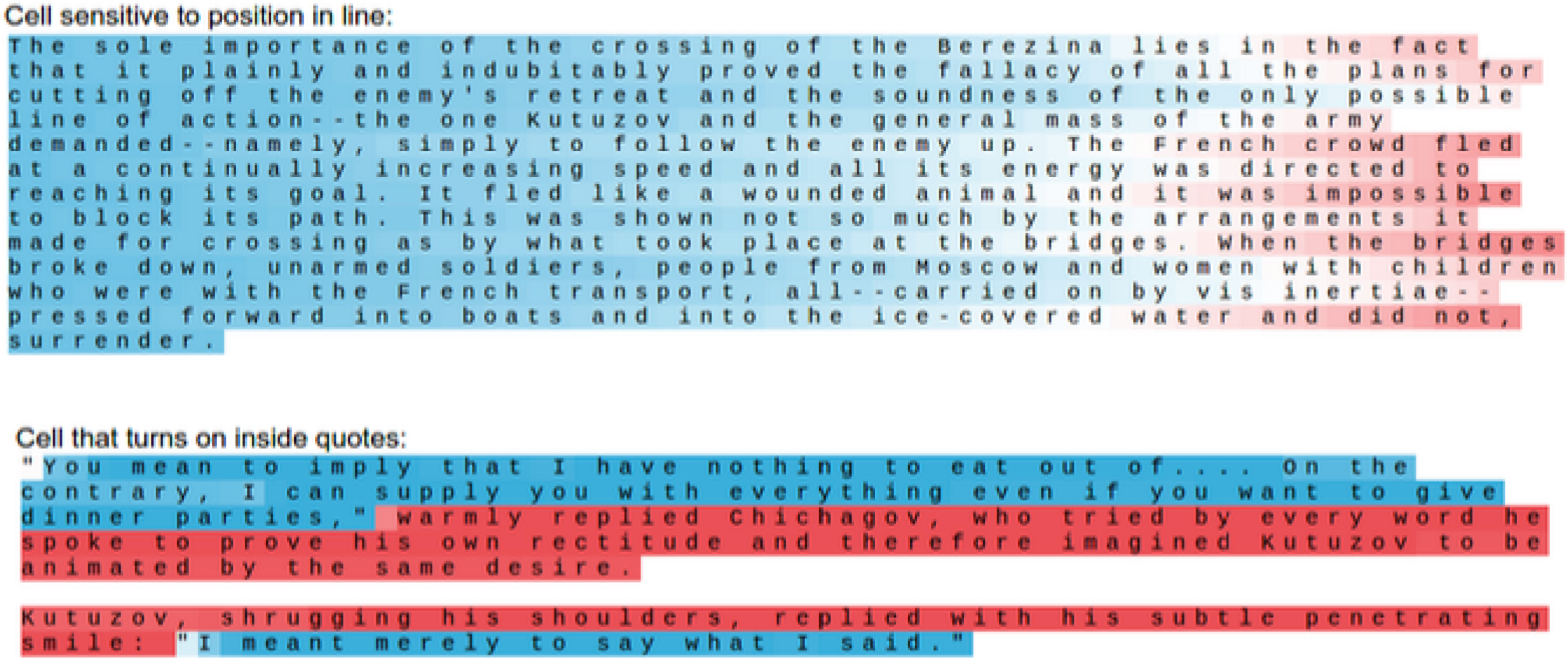
Why did the network reach this decision?

- Tracing back decisions to inputs

⇒ Causal explanation

identifying neurons

Visualizing Individual Cells



Cell sensitive to position in line:

The sole importance of the crossing of the Berezina lies in the fact
that it plainly and indubitably proved the fallacy of all the plans for
cutting off the enemy's retreat and the soundness of the only possible
line of action--the one Kutuzov and the general mass of the army
demanded--namely, simply to follow the enemy up. The French crowd fled
at a continually increasing speed and all its energy was directed to
reaching its goal. It fled like a wounded animal and it was impossible
to block its path. This was shown not so much by the arrangements it
made for crossing as by what took place at the bridges. When the bridges
broke down, unarmed soldiers, people from Moscow and women with children
who were with the French transport, all--carried on by vis inertiae--
pressed forward into boats and into the ice-covered water and did not,
surrender.

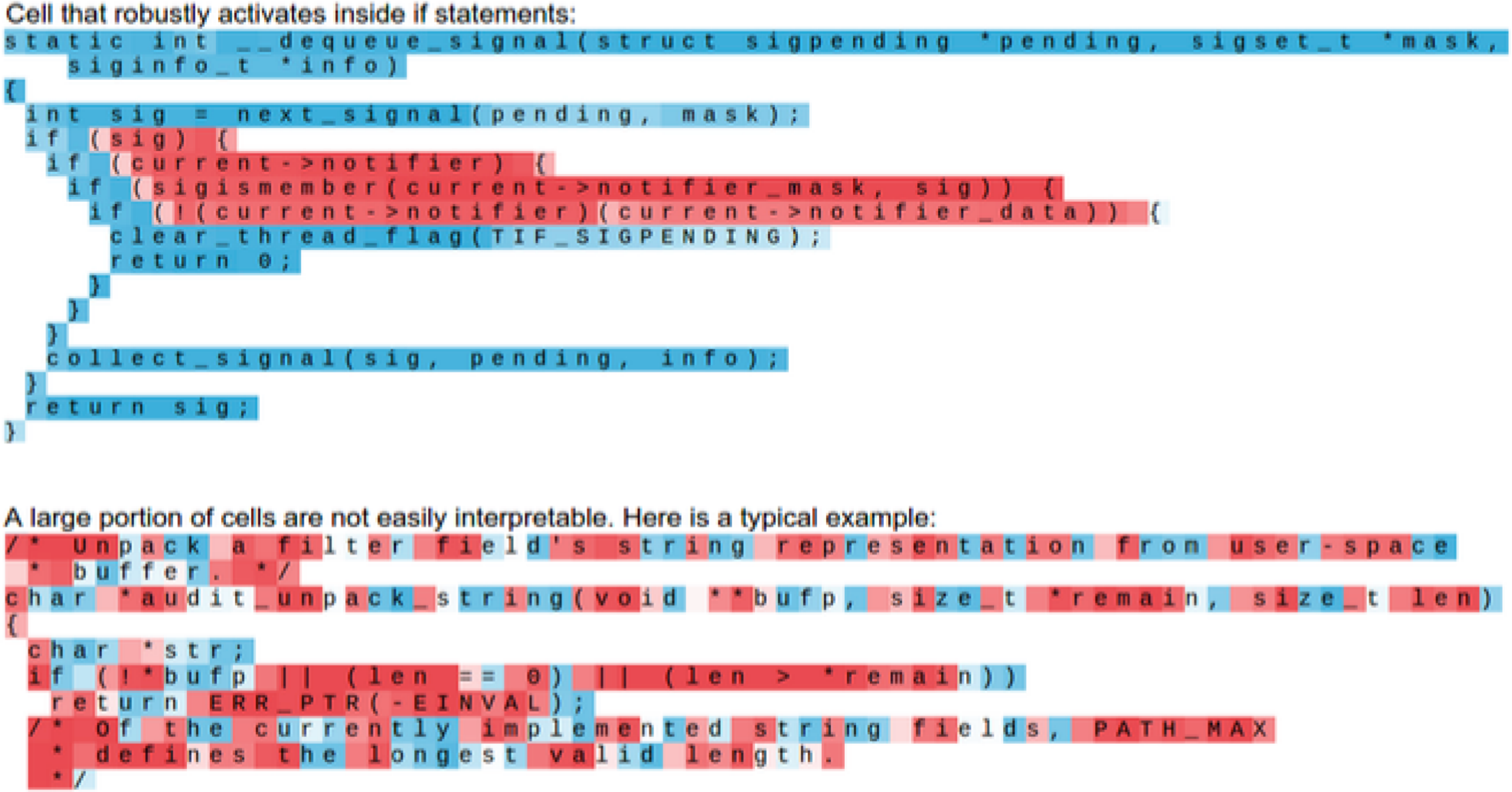
Cell that turns on inside quotes:

"You mean to imply that I have nothing to eat out of.... On the
contrary, I can supply you with everything even if you want to give
dinner parties," warmly replied Chichagov, who tried by every word he
spoke to prove his own rectitude and therefore imagined Kutuzov to be
animated by the same desire.

Kutuzov, shrugging his shoulders, replied with his subtle penetrating
smile: "I meant merely to say what I said."

Karpathy et al. (2015): "Visualizing and Understanding Recurrent Networks"

Visualizing Individual Cells



Cell that robustly activates inside if statements:

```
static int __dequeue_signal(struct sigpending *pending, sigset_t *mask,
   siginfo_t *info)
{
 int sig = next_signal(pending, mask);
 if (sig) {
  if (current->notifier) {
   if (sigismember(current->notifier_mask, sig)) {
    if (!(current->notifier)(current->notifier_data)) {
     clear_thread_flag(TIF_SIGPENDING);
     return 0;
    }
   }
  }
  collect_signal(sig, pending, info);
 }
 return sig;
}
```

A large portion of cells are not easily interpretable. Here is a typical example:

```
/* Unpack a filter field's string representation from user-space
 * buffer. */
char *audit_unpack_string(void **bufp, size_t *remain, size_t len)
{
 char *str;
 if (!*bufp || (len == 0) || (len > *remain))
  return ERR_PTR(-EINVAL);
 /* Of the currently implemented string fields, PATH_MAX
  * defines the longest valid length.
  */
```

Identifying Neurons

- How are specific properties encoded?
- Easiest case: in a single neuron
- How do we find it?
- Example: length of sequence
  - given: encoder-decoder model without attention
  - does the encoder record the length of the consumed sequence?
  - does the decoder record the length of the generated sequence?

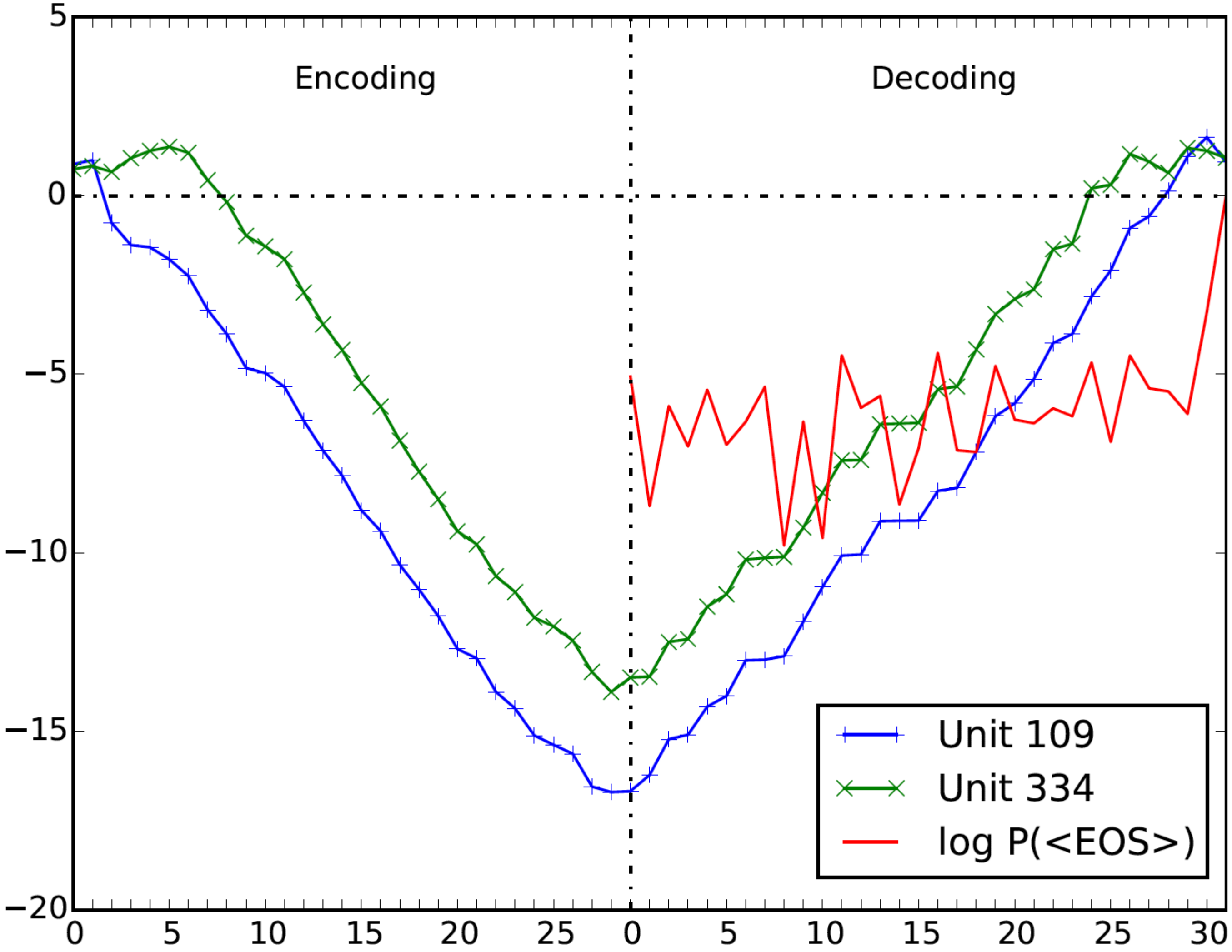
Correlation

- Select a neuron

- Compute correlation
  - value of neuron when processing xth word
  - position x

- Success if highly correlated neuron found

Neurons Correlated with Length



"Why neural translations are the right length" (Shi, Knight, Yuret, EMNLP 2016) questions?