NUS National University of Singapore | School of Computing

Sequence Applications

CS4248 Natural Language Processing

Week 12

Min-Yen KAN

12

Click to edit Master Attribution style.

Recap of Week 11

Many classification tasks becomes supervised machine learning

- Sentiment Analysis, Summarization and Question Answering
- Among many others…

They can be accompanied by the definition of good feature classes (rather than individual features)

Manipulate natural language to engineer features and lexicons for use in tasks

Click to edit Master Attribution style.

Week 12 Agenda

Contextual Word Embeddings

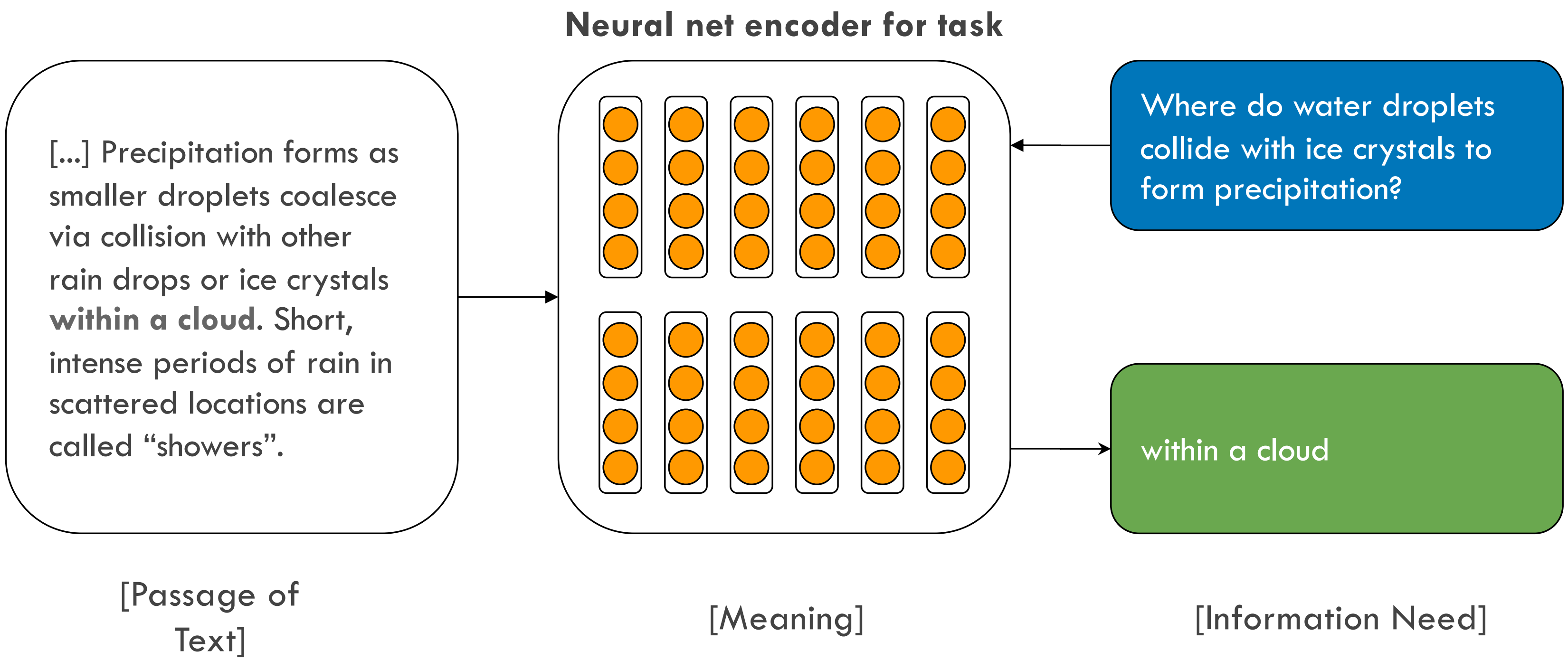
Machine Translation

Question Answering II

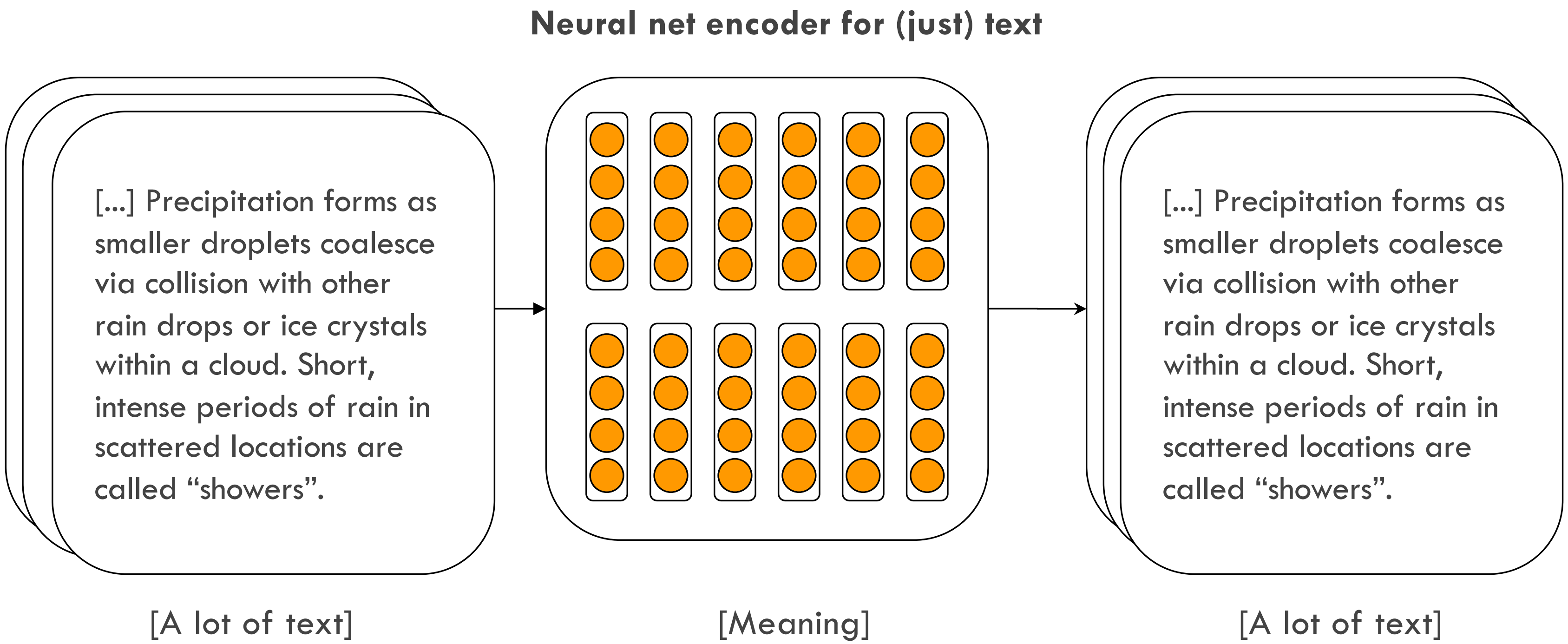
Contextual Word Embeddings

Revisiting Word Embeddings with Seq2Seq

Supervised training

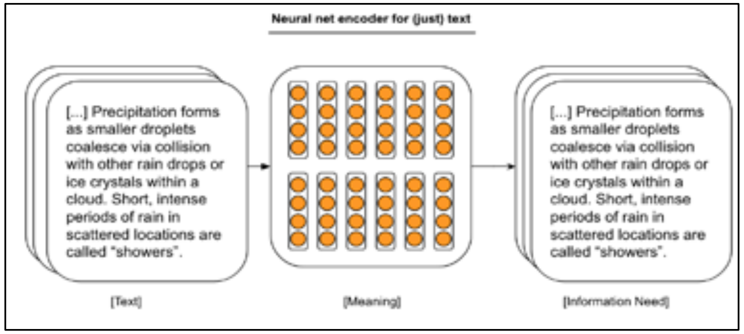


Unsupervised pretrained representations



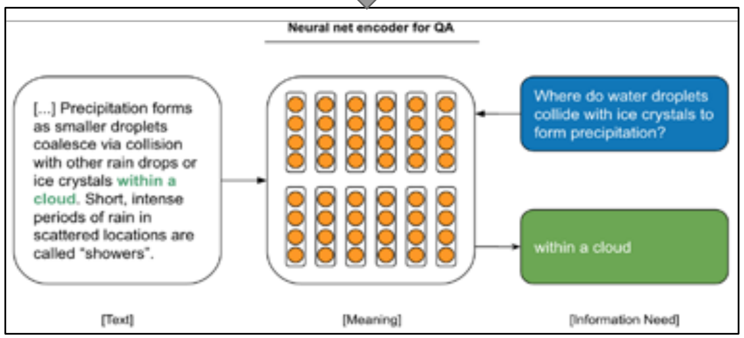
Lifting over pretrained representations

Pretrained Language Model



Transfer

Task (i.e., Machine Reading)



How is it different from pretrained word embeddings?

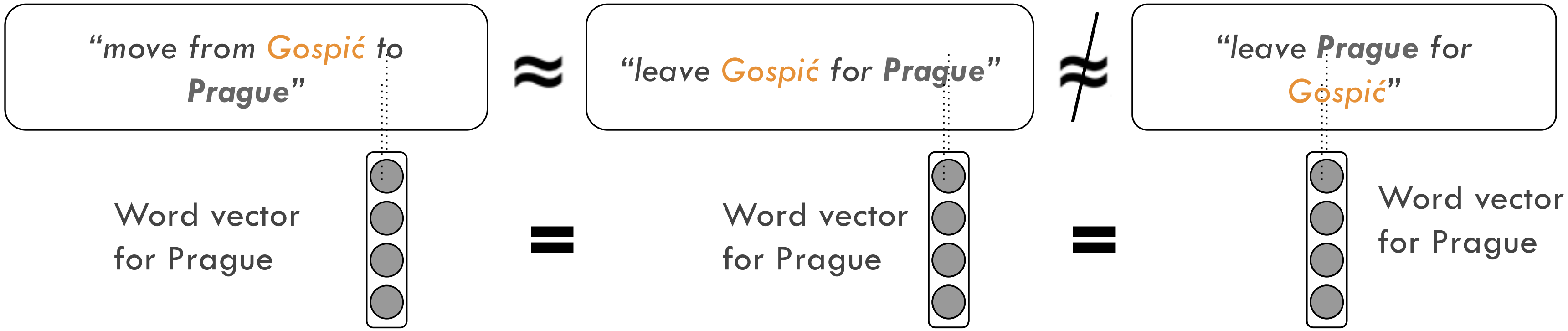
Pretrained Word Embeddings (word2vec)

- Predicting co-occurrence of words
- Independent of other context

Pretrained Contextualized Embeddings (e.g. ELMo, BERT)

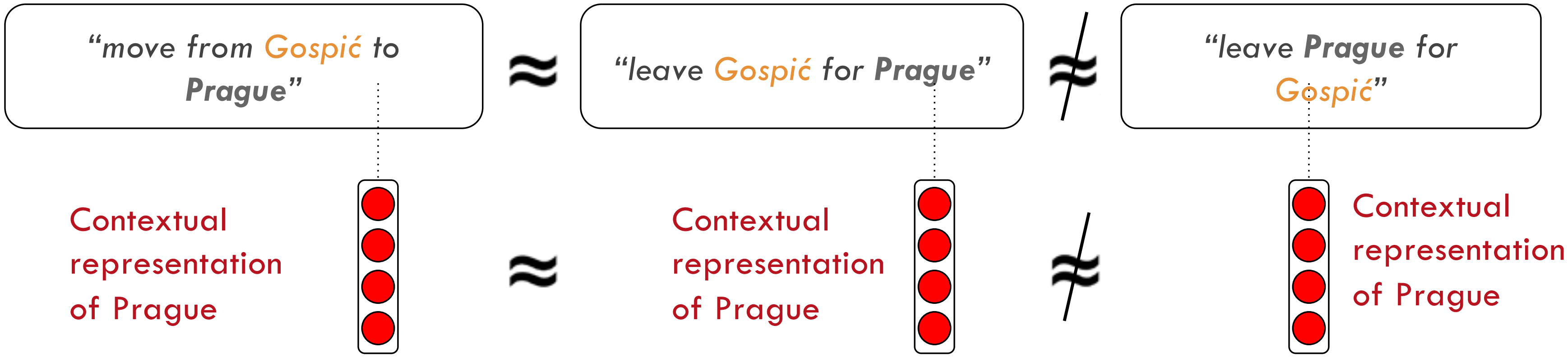
- Predicting whole text (using LSTM, or Self-Attention)
- Full dependence on other context

Representing Words in Context



Word representations should vary depending on context.

Representing Words in Context



Word representations should vary depending on context.

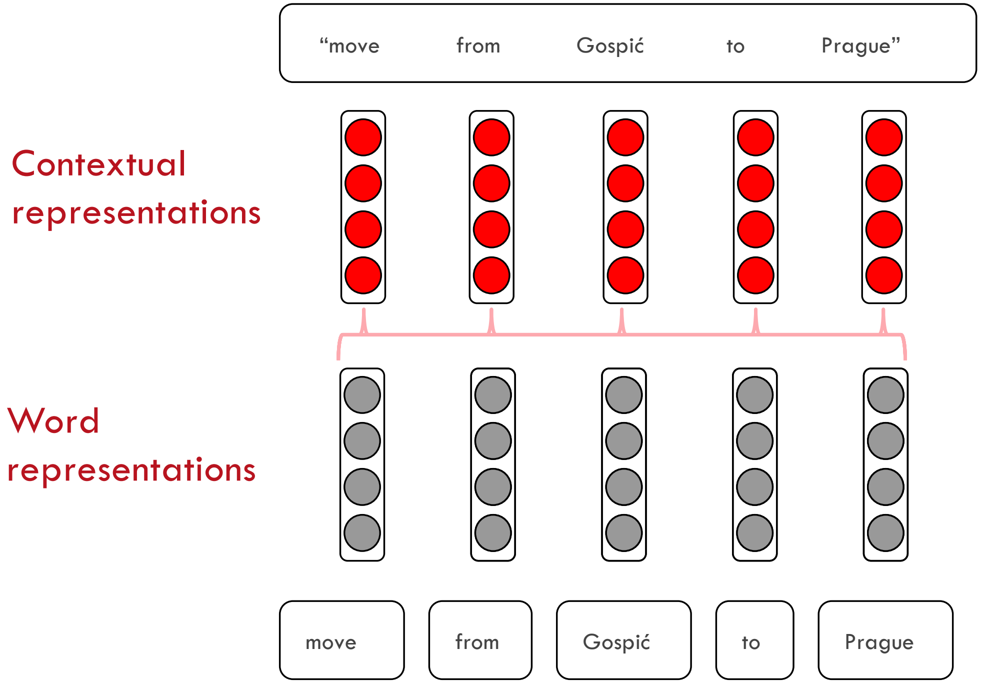
Contextual word representation:

- a word representation, computed *conditionally* on the given context

Representing Words in Context

Composition of word vectors into contextualized word representations

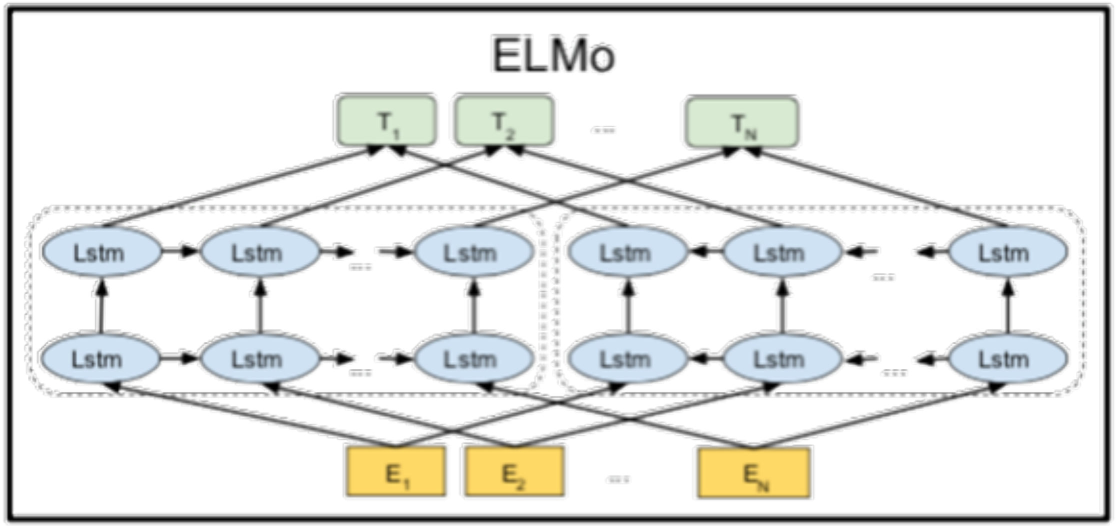
Idea: Use vector composition function



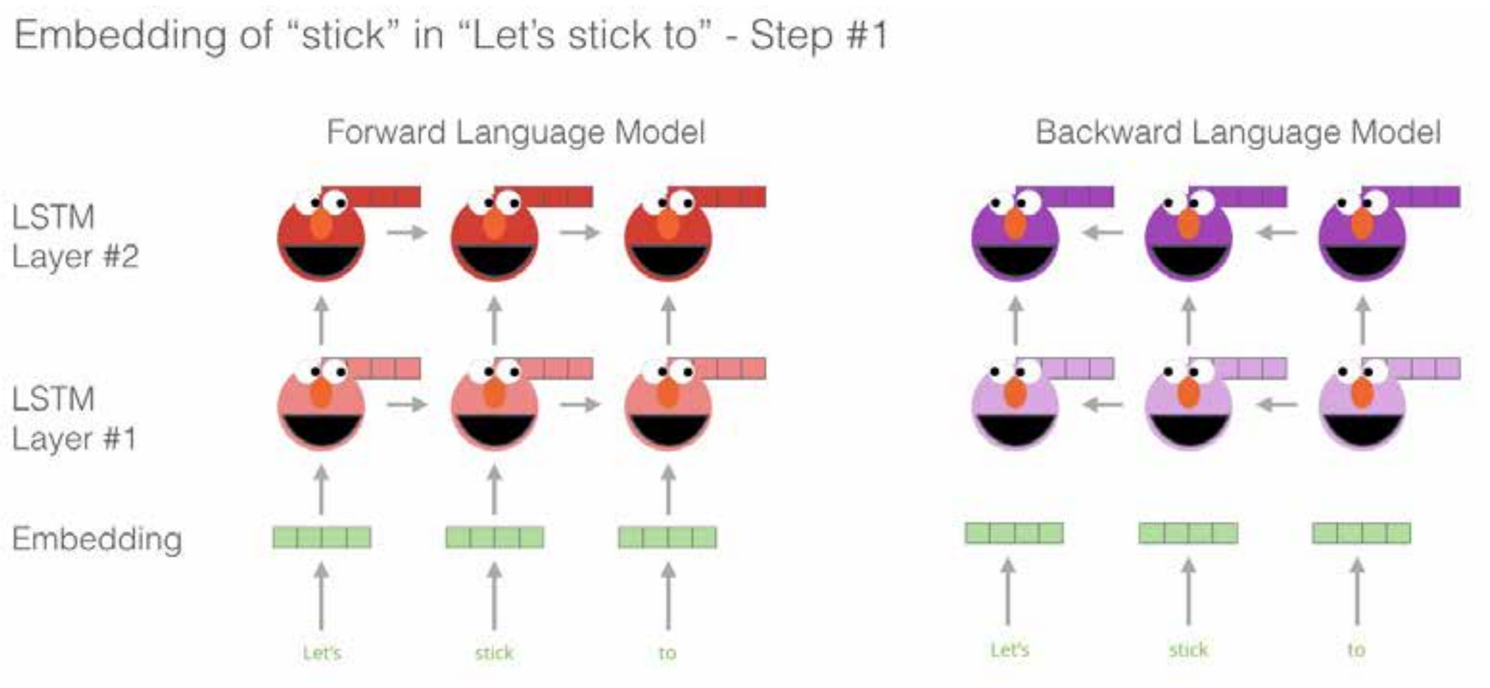
ELMo: **E**mbeddings from **L**anguage **Mo**dels

Peters et al. (2018) [Deep Contextualized Word Representations]()

- Train a BiLSTM for Bidirectional language modeling on a large dataset
- Encode the sentence bidirectionally through both forward and backward LSTMs
- Combine both representations into final contextual embeddings



Embeddings from Language Models

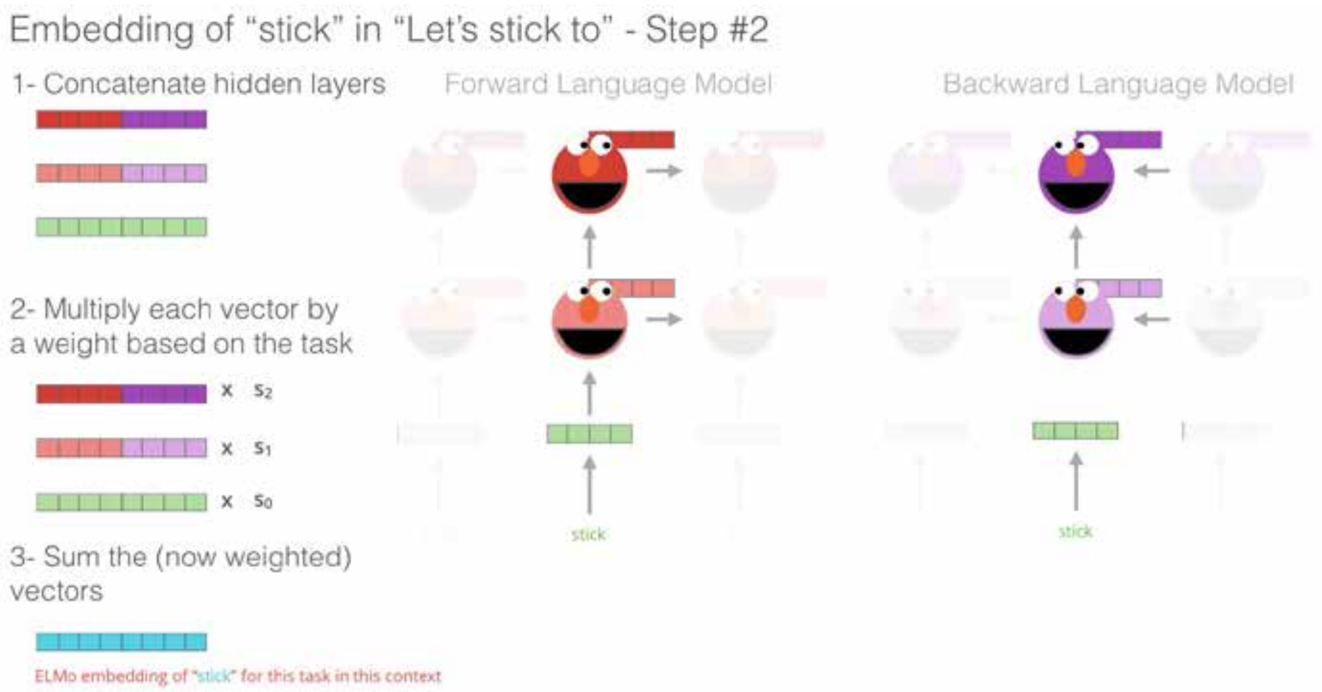


Figures from http://jalammar.github.io/illustrated-bert/

Embeddings from Language Models

Task-specific weighting!

Train end-to-end with task



Figures from http://jalammar.github.io/illustrated-bert/

CWE significantly augment performance

	Task	Previous SOTA		Our baseline	ELMo + baseline	Increase (absolute/ relative)
Machine Reading -	SQuAD	Liu et al. (2017)	84.4	81.1	85.8	4.7 / 24.9%
Textual Entailment -	SNLI	Chen et al. (2017)	88.6	88.0	88.7 $\pm$ 0.17	0.7 / 5.8%
Semantic Labeling -	SRL	He et al. (2017)	81.7	81.4	84.6	3.2 / 17.2%
Coreference Resolution -	Coref	Lee et al. (2017)	67.2	67.2	70.4	3.2 / 9.8%
Entity Extraction -	NER	Peters et al. (2017)	91.93 $\pm$ 0.19	90.15	92.22 $\pm$ 0.10	2.06 / 21%
Sentiment Analysis -	SST-5	McCann et al. (2017)	53.7	51.4	54.7 $\pm$ 0.5	3.3 / 6.8%

What does ELMo learn?

Disambiguating the meaning of words in context

- POS, word sense, etc.

	Source	Nearest Neighbors
GloVe	play	playing, game, games, played, players, plays, player, Play, football, multiplayer
biLM	Chico Ruiz made a spectacular play on Alusik 's grounder {. . . }	Kieffer , the only junior in the group , was commended for his ability to hit in the clutch , as well as his all-round excellent play .
	Olivia De Havilland signed to do a Broadway play for Garson {. . . }	{. . . } they were actors who had been handed fat roles in a successful play , and had talent enough to fill the roles competently , with nice understatement .

Noisy Channel Model

Viewing translation as denoising

MT as code breaking

One naturally wonders if the problem of translation could conceivably be treated as a problem in cryptography. When I look at an article in Russian, I say: '***This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode.***'

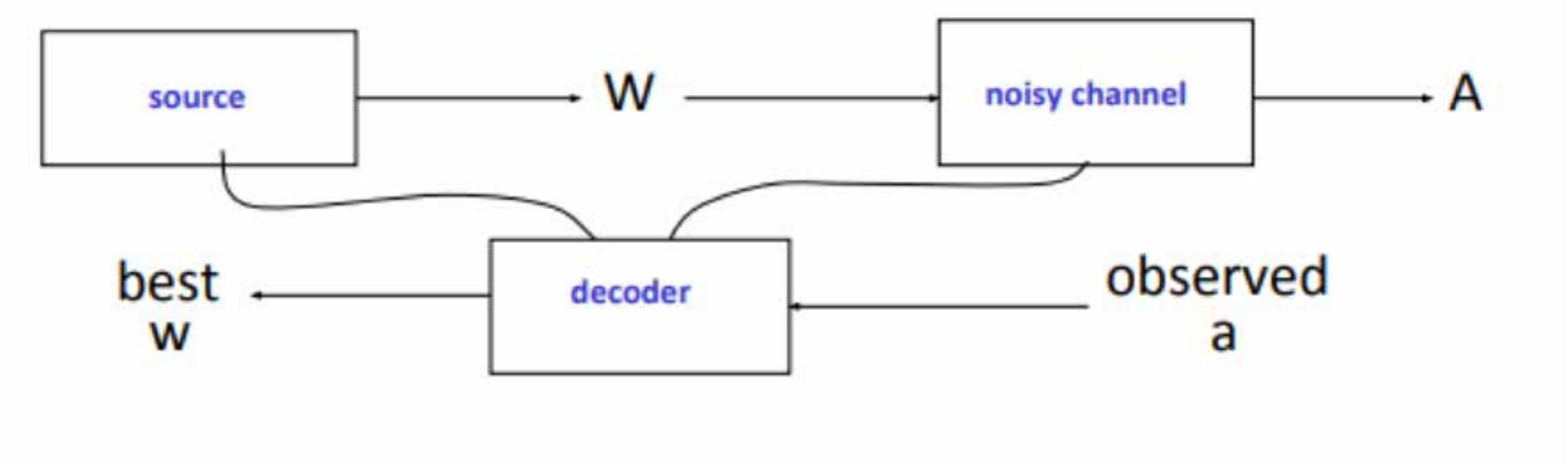
Warren Weaver to Norbert Wiener, March, 1947

Slide Credits: Diyi Yang (Georgia Tech)

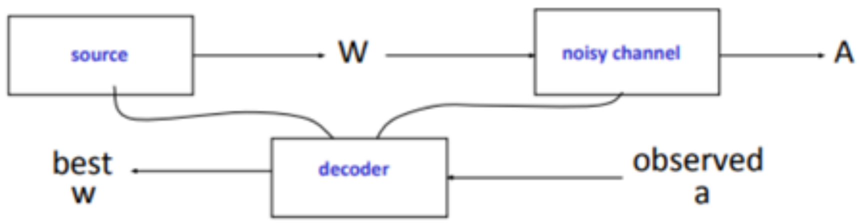
The Noisy Channel Model

Slide Credits: Diyi Yang (Georgia Tech)

The Noisy Channel Model



Slide Credits: Diyi Yang (Georgia Tech)



We want to predict a sentence given acoustics:

$$\hat{w} = \arg\max P(w|a)$$

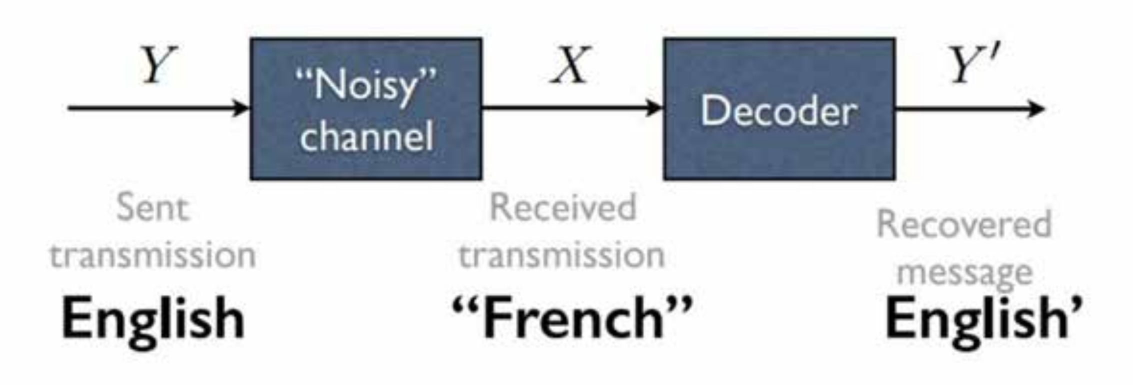
$$= \arg\max P(a|w)\,P(w)\,/\,P(a)$$ Bayes' Rule

$$= \arg\max P(a|w)\,P(w)$$ Input (the observed sound) is constant

Slide Credits: Diyi Yang (Georgia Tech)

Noisy Channel MT

We can apply this Idea to MT.



$$\hat{e} = \arg\max_e P_{LM}(e) \times P_\theta(f|e)$$

Slide Credits: Diyi Yang (Georgia Tech)

MT as Direct Modeling

$$\hat{e} = \arg\max_{e} P_{\theta}(e|f)$$

One model does everything

Trained to reproduce a corpus of translations

Slide Credits: Diyi Yang (Georgia Tech)

Two Views of MT

Code breaking (aka the noisy channel, Bayes rule)

- I know the target language
- I have example translations texts (exam enciphered data)
- Statistical Machine Translation (SMT)

Direct modeling (aka pattern matching)

- I have really good learning algorithms and a bunch of example inputs (source language sentences) and outputs (target language translations)
- Neural Machine Translation (NMT)

Slide Credits: Diyi Yang (Georgia Tech)

Which is better?

Noisy Channel: $\hat{e} = \arg\max_{e} P_{LM}(e) \times P_{\theta}(f|e)$

- Can leverage monolingual target language data
- Search happens under a product of two models (individual models can be simple, product can be powerful)

Direct Model: $\hat{e} = \arg\max_{e} P_{\theta}(e|f)$

- Directly model the process you care about
- Model must be very powerful

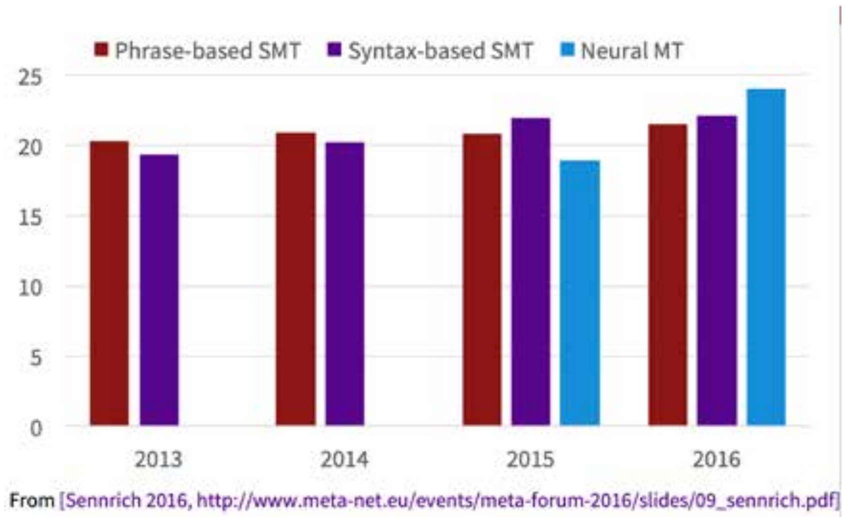
Slide Credits: Diyi Yang (Georgia Tech)

Where are we now?

Direct modeling is where most of the action is

- Neural networks are very good at generalizing and conceptually very simple
- Inference in "product of two models" is hard

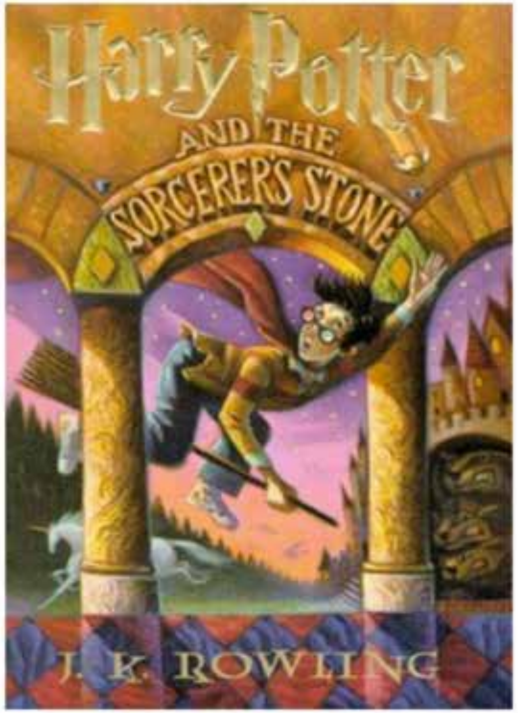
But noisy channel ideas are incredibly important and still play a big role in how we think about translation

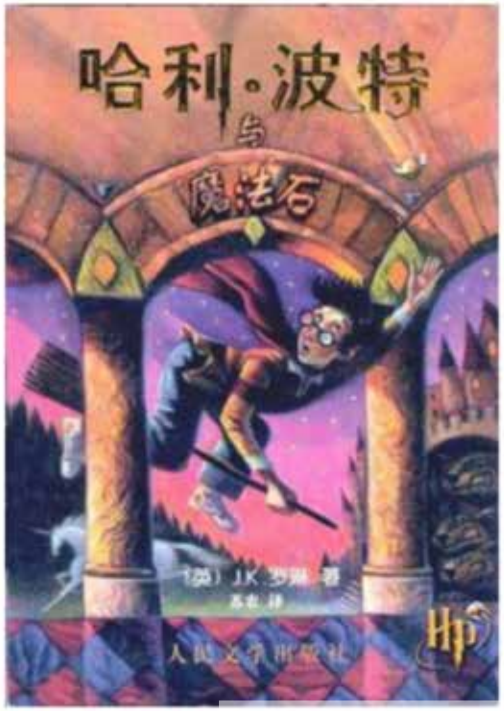


Slide Credits: Diyi Yang (Georgia Tech)

Parallel Corpora

Parallel Corpora





Slide Credits: Diyi Yang (Georgia Tech)


DANGER–KEEP OUT !
危險，請避開！
BAHAYA–JANGAN DEKAT !
அபாயம்-அருகில்வராதீர்கள்!

Self translation

Ling et al. (2013) Mining Parallel Corpora From Sina Weibo and Twitter

	ENGLISH	MANDARIN
1	i **wanna** live in a wes anderson world	我想要生活在Wes Anderson的世界里
2	Chicken soup, corn never truly digests. **TMI.**	鸡汤吧，玉米神马的从来没有真正消化过.恶心
3	To DanielVeuleman **yea iknw imma** work on that	对DanielVeuleman说，是的我知道，我正在向那方面努力
4	**msg 4** Warren G his **cday** is today 1 **yr** older.	发信息给Warren G，今天是他的生日，又老了一岁了。
5	Where **the hell** have you been all these years?	这些年你**TMD**到哪去了
	ENGLISH	**ARABIC**
6	It's **gonna** be a warm week!	الاسبو ع **الياي** حر
7	onni this gift only **4 u**	أوني هذة الهدية فقط لك
8	sunset in aqaba **:)**	غروب الشمس في العقبة:)
9	RT @MARYAMALKHAWAJA: there is a call for widespread protests in #bahrain **tmrw**	هناك نداء لظاهرات في عدة مناطق غدا

Table 2: Examples of English-Mandarin and English-Arabic sentence pairs. The English-Mandarin sentences were extracted from Sina Weibo and the English-Arabic sentences were extracted from Twitter. Some messages have been shorted to fit into the table. Some interesting aspects of these sentence pairs are marked in bold.

Slide Credits: Diyi Yang (Georgia Tech)

But also comparable corpora

Distant or weak supervision or heuristics to find *almost* parallel corpora.

More monolingual data

There is a lot more monolingual data in the world than translated data

Easy to get about 1 trillion words of English by crawling the web

With some work, you can get 1 billion translated words of English–French

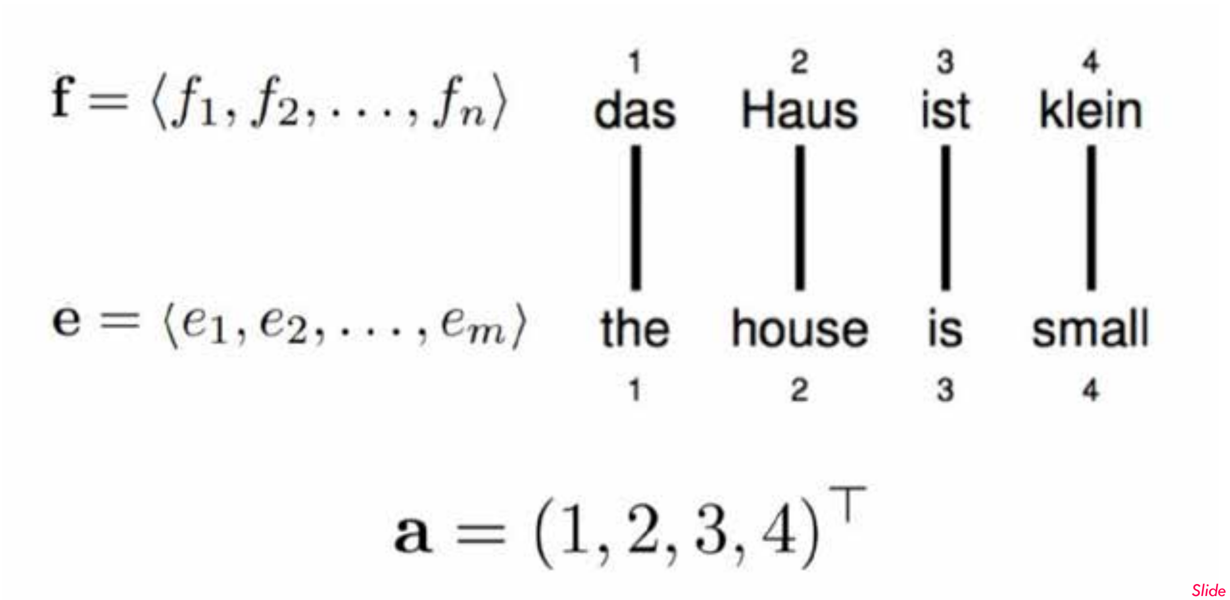
- But what about Japanese–Turkish?

Slide Credits: Diyi Yang (Georgia Tech)

Word Alignment

Word Alignment

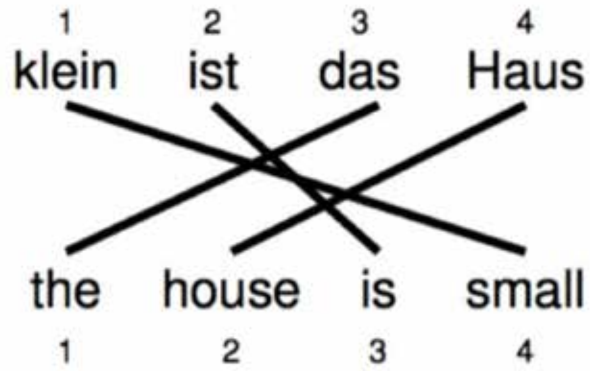
Alignment can be visualized by drawing links between two sentences, and they are represented as vectors of positions



Slide Credits: Diyi Yang (Georgia Tech)

Reordering

Words can be **reordered** when translated



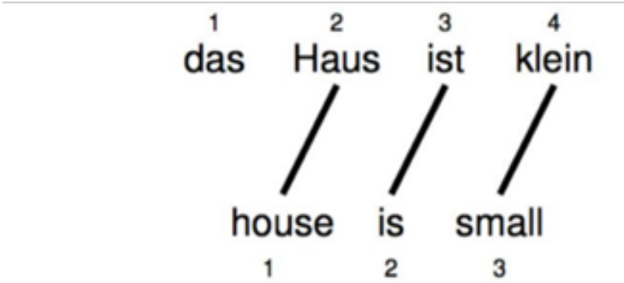
$$\mathbf{a} = (3, 4, 2, 1)^\top$$

Slide Credits: Diyi Yang (Georgia Tech)

Word Dropping

Words can be reordered, **dropped** when translated

A source word may not be translated at all



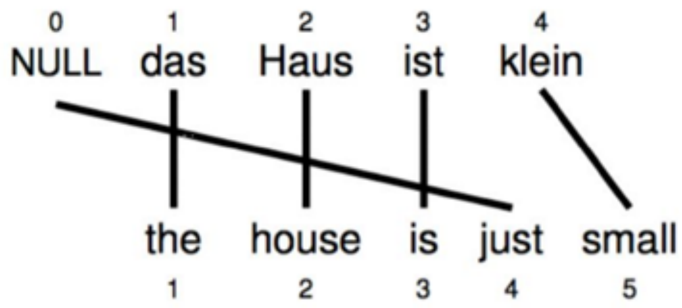
$$\mathbf{a} = (2, 3, 4)^\top$$

Slide Credits: Diyi Yang (Georgia Tech)

Word Insertion

Words can be reordered, dropped, **inserted** during translation

- English *just* does not have an equivalent
- But it must be explained – we typically assume every source sentence contains a NULL token



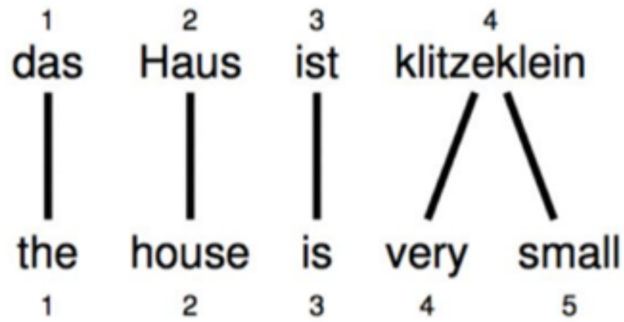
$$\mathbf{a} = (1, 2, 3, 0, 4)^\top$$

Slide Credits: Diyi Yang (Georgia Tech)

Word Fertility: one-to-many

Words can be reordered, dropped, inserted, **multiply translated** during translation

A source word may translate into more than one target word



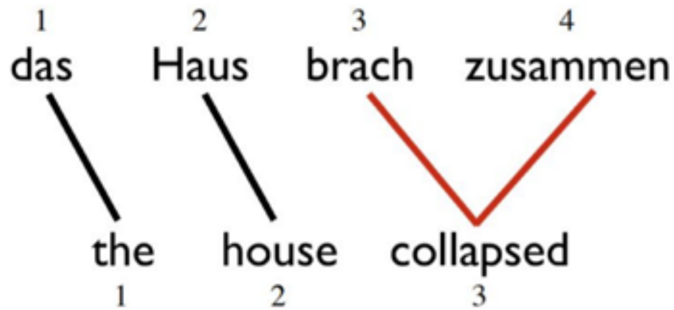
$$\mathbf{a} = (1, 2, 3, 4, 4)^\top$$

Slide Credits: Diyi Yang (Georgia Tech)

Many-to-one translation

Words can be reordered, dropped, inserted, **multiply translated (in both senses)** during translation.

- More than one source word may not translate as a unit in lexical translation



$\mathbf{a} = ???$ $\qquad$ $\mathbf{a} = (1, 2, (3, 4)^\top)^\top$?

Slide Credits: Diyi Yang (Georgia Tech)

Computing Word Alignments

Word alignments are the basis for most translation algorithms

Given two sentences F and E, find a good alignment

But a word-alignment algorithm can also be part of a mini-translation model itself

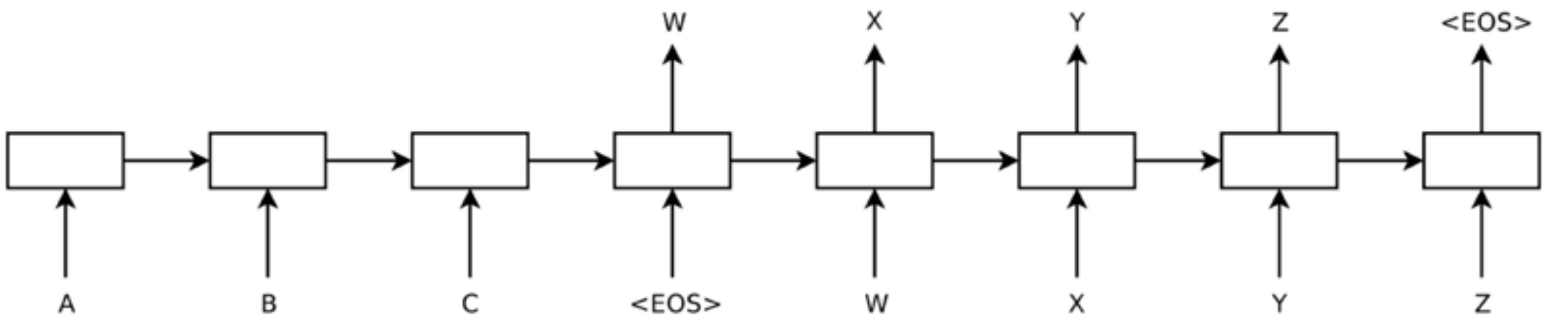
One the most basic alignment models is also a simplistic translation model

Slide Credits: Diyi Yang (Georgia Tech)

Sentence Encoding

A Bottleneck in representation

Conditional LM: Encoder–Decoder



Slide Credits: Diyi Yang (Georgia Tech)

Neural Machine Translation

The probability of translation y given the source sentence x

$$\log p(y|x) = \sum_{j=1}^{m} \log p\left(y_j | y_{<j}, \boldsymbol{s}\right)$$

Encoded vector generated from the sequence of hidden states

where

$$p\left(y_j | y_{<j}, \boldsymbol{s}\right) = \operatorname{softmax}\left(g\left(\boldsymbol{h}_j\right)\right) \qquad \boldsymbol{h}_j = f(\boldsymbol{h}_{j-1}, \boldsymbol{s}),$$

Slide Credits: Diyi Yang (Georgia Tech)

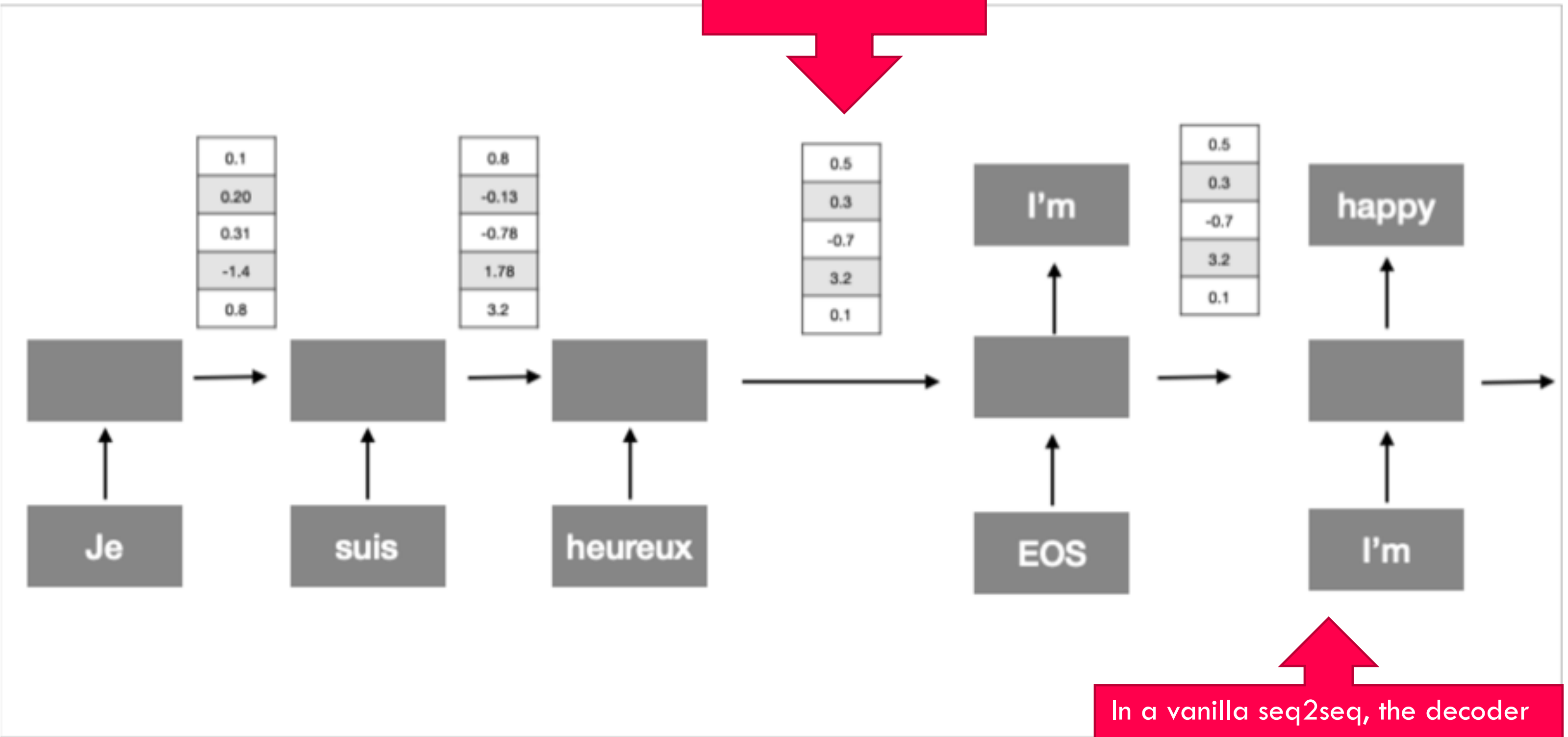
Training Objective

$$L_t = \sum\nolimits_{(x,y)} - \log P(y|x)$$

As in other RNNs, we can train by minimizing the loss between what we predict at each time step and the ground truth.

Slide Credits: Diyi Yang (Georgia Tech)

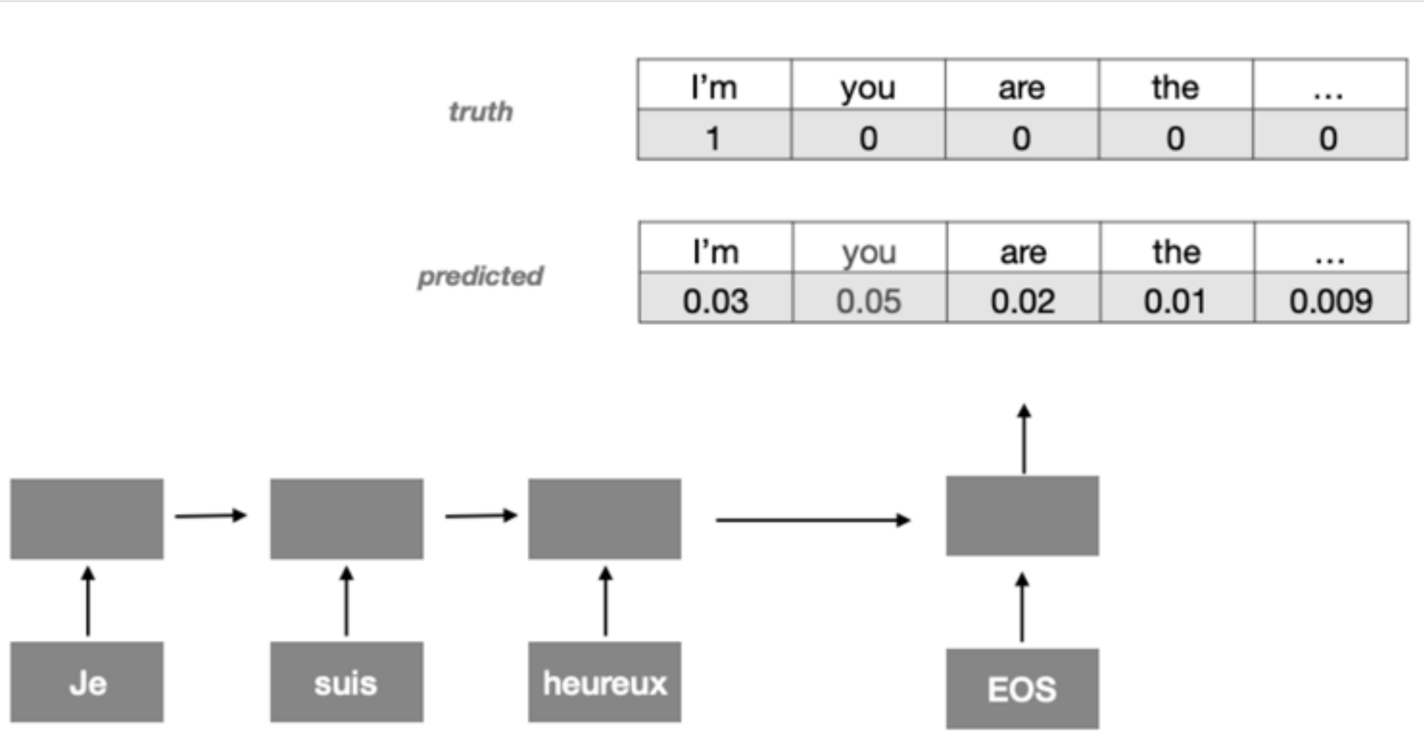
The entire input is summarized in this one single vector



In a vanilla seq2seq, the decoder state depends just on the previous state and the previous output

Slide Credits: Diyi Yang (Georgia Tech)

Incorrect Translation



Slide Credits: Diyi Yang (Georgia Tech)

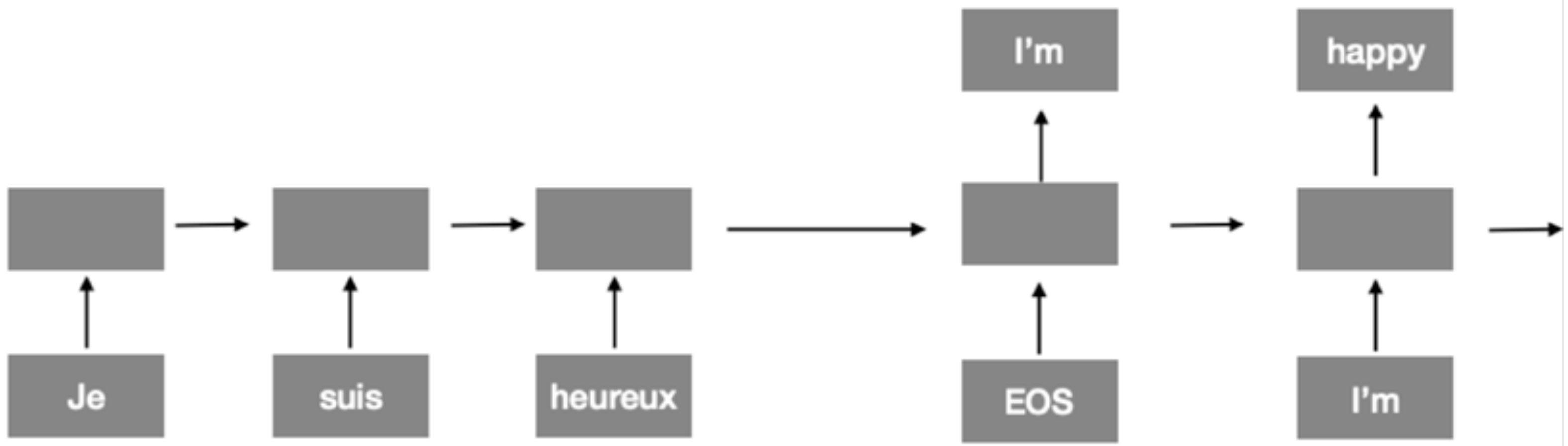
Correct, but still needs tuning

truth

happy	great	bad	ok	...
1	0	0	0	0

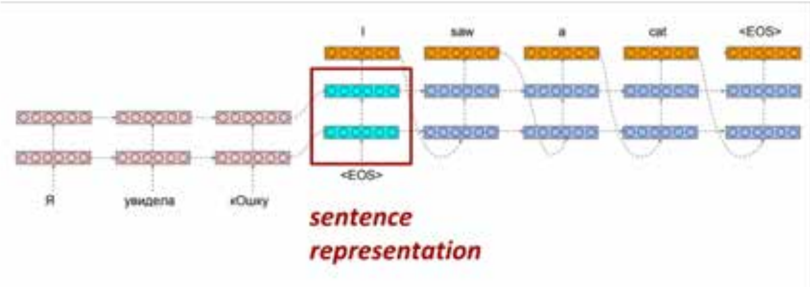
predicted

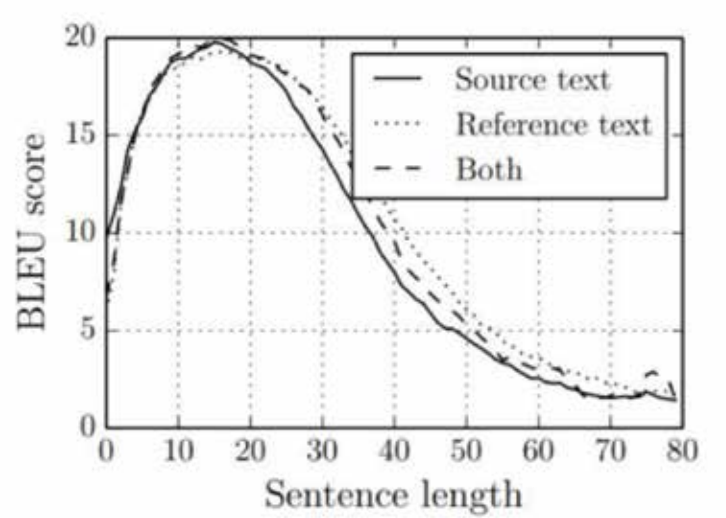
happy	great	bad	ok	...
0.13	0.08	0.01	0.03	0.009



Slide Credits: Diyi Yang (Georgia Tech)

Representation Bottleneck





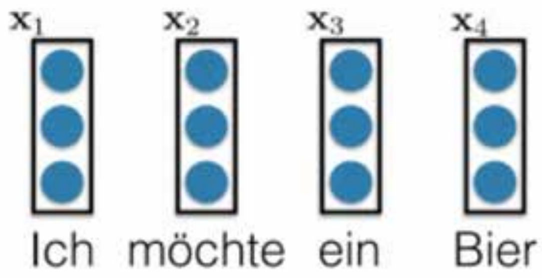
- Fixed sized representation degrades as sentence length increases
- Compressing the entire input sentence into a vector basically says "memorize the sentence"
- Common sense experience says translators refer back and forth to the input (also backed up by eye-tracking studies)

Slide Credits: Diyi Yang (Georgia Tech)

Encoder–Decoder with Attention

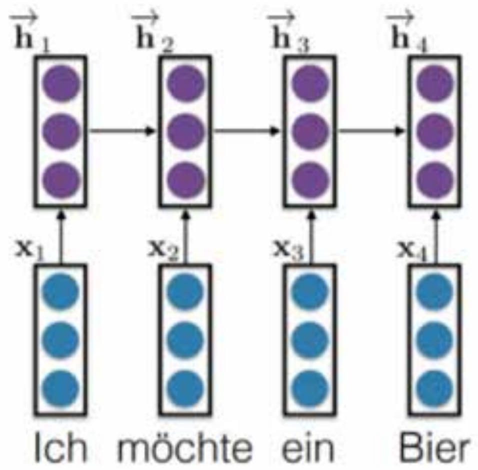
A standard NMT model

Encoder: Bidirectional RNN



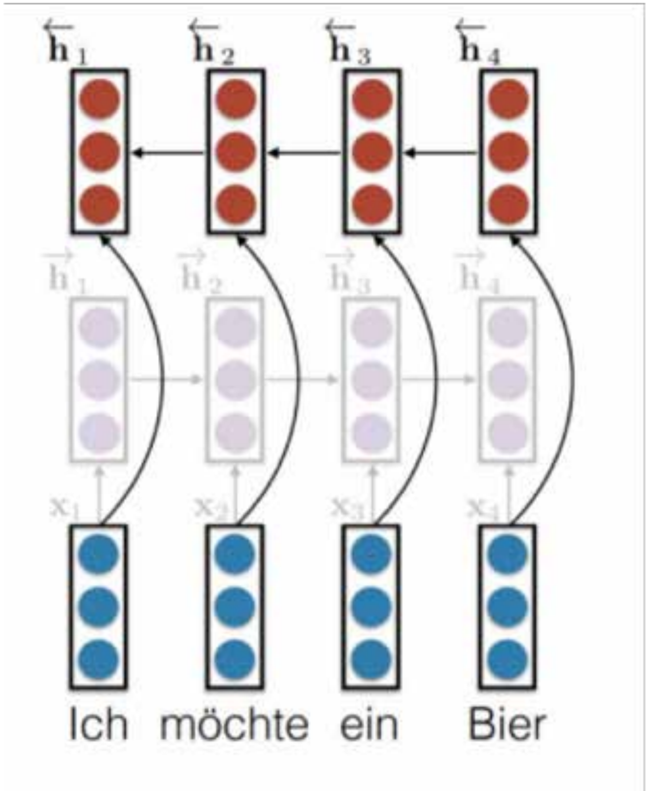
Slide Credits: Diyi Yang (Georgia Tech)

Encode Forwards



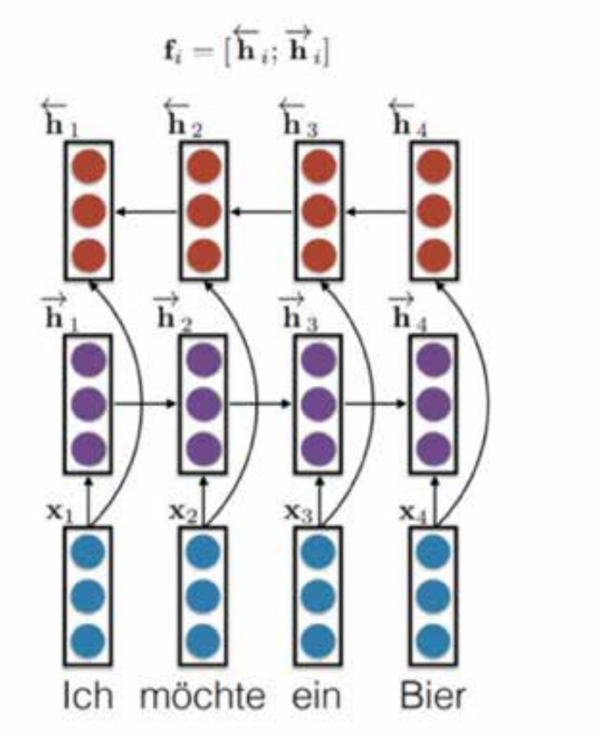
Slide Credits: Diyi Yang (Georgia Tech)

Encode Backwards

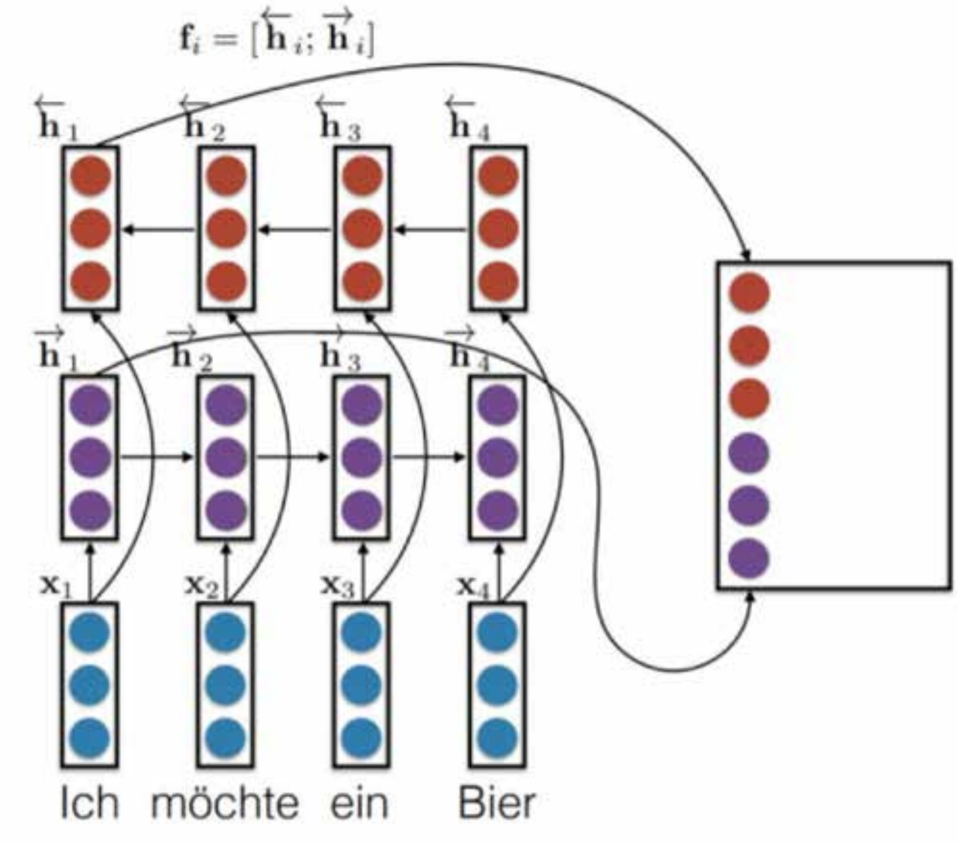


Slide Credits: Diyi Yang (Georgia Tech)

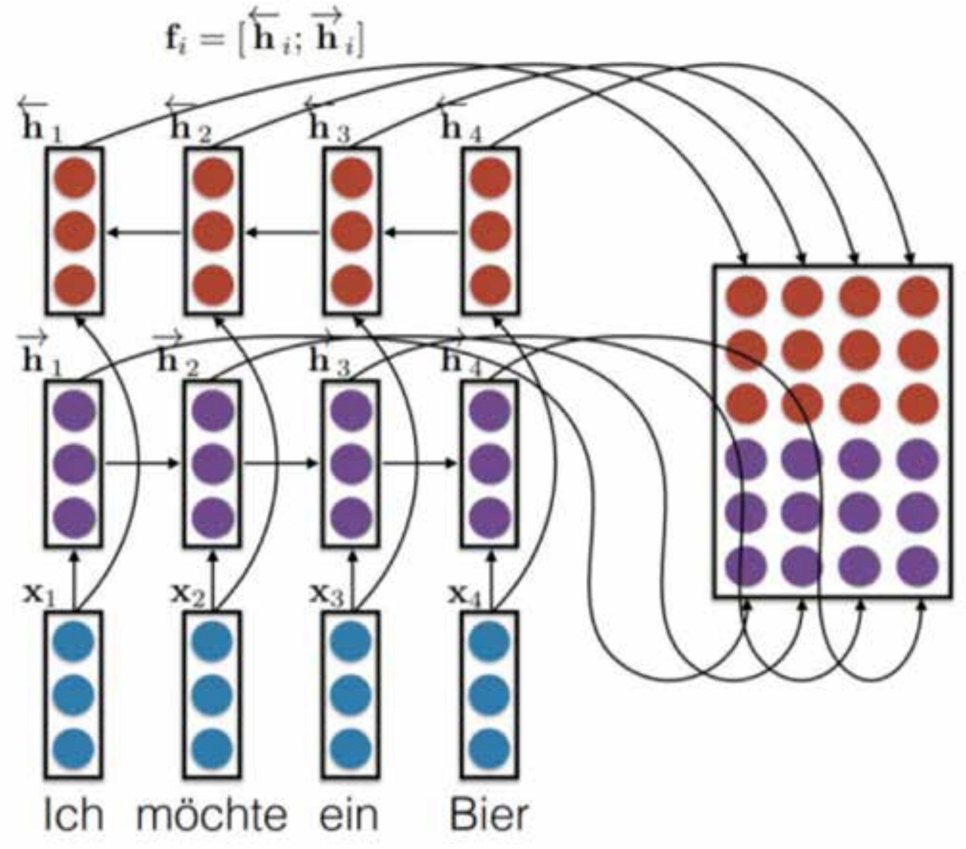
Concatenate



Slide Credits: Diyi Yang (Georgia Tech)

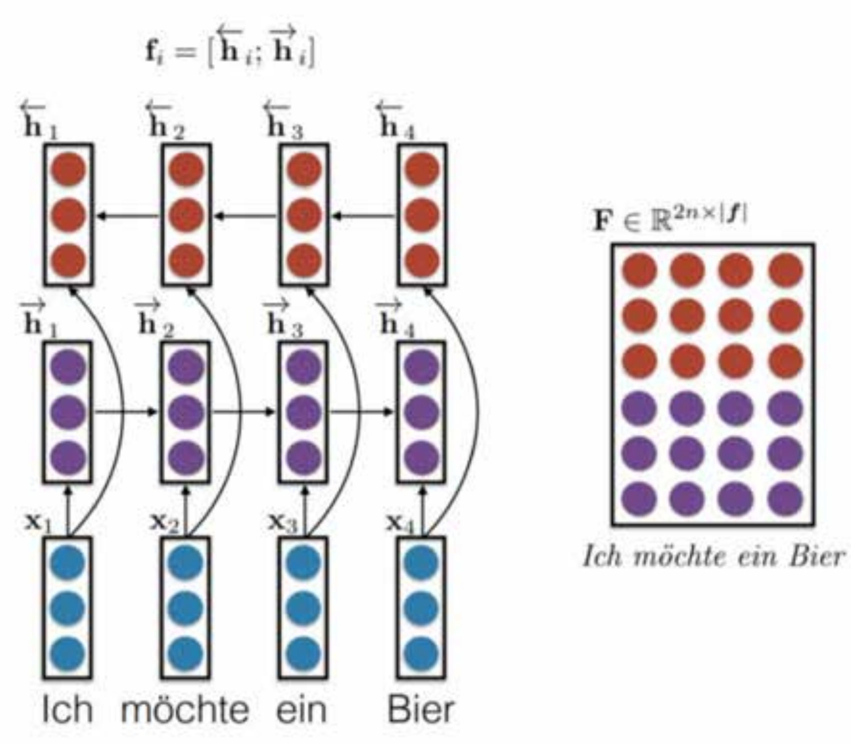


Slide Credits: Diyi Yang (Georgia Tech)



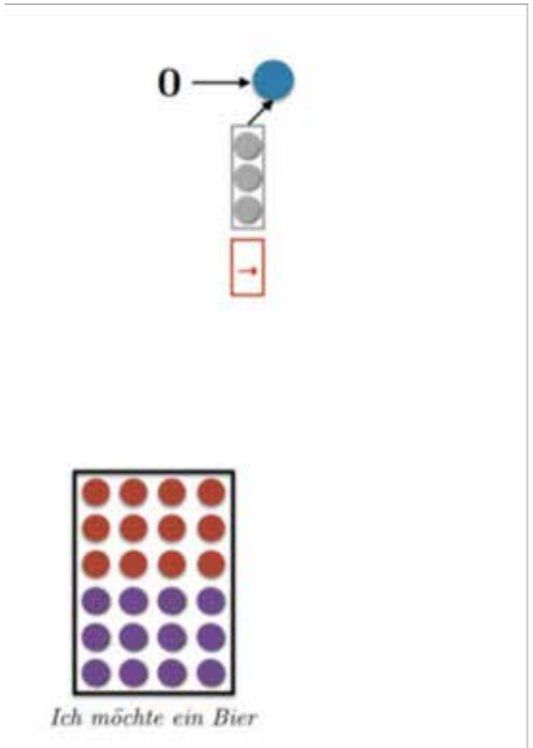
Slide Credits: Diyi Yang (Georgia Tech)

Matrix Sentence Encoding

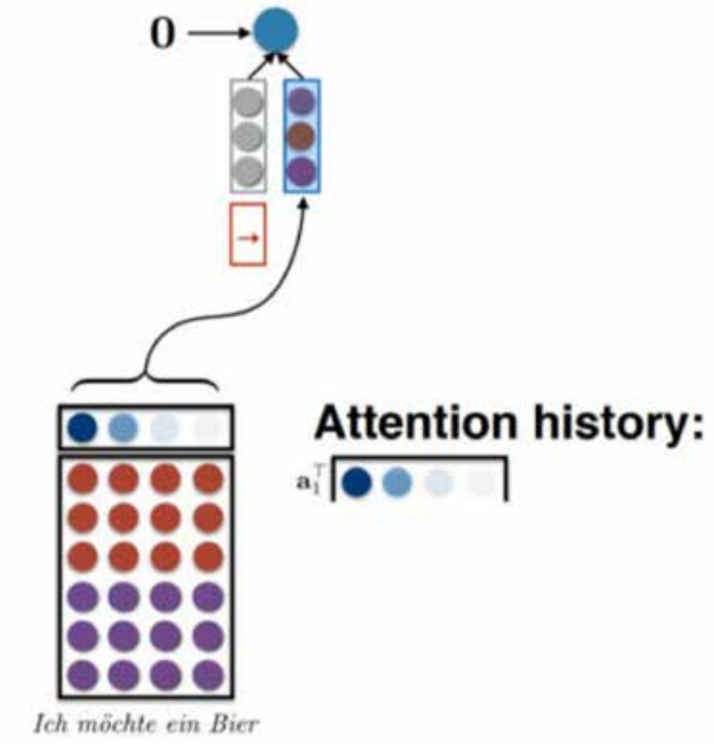


Slide Credits: Diyi Yang (Georgia Tech)

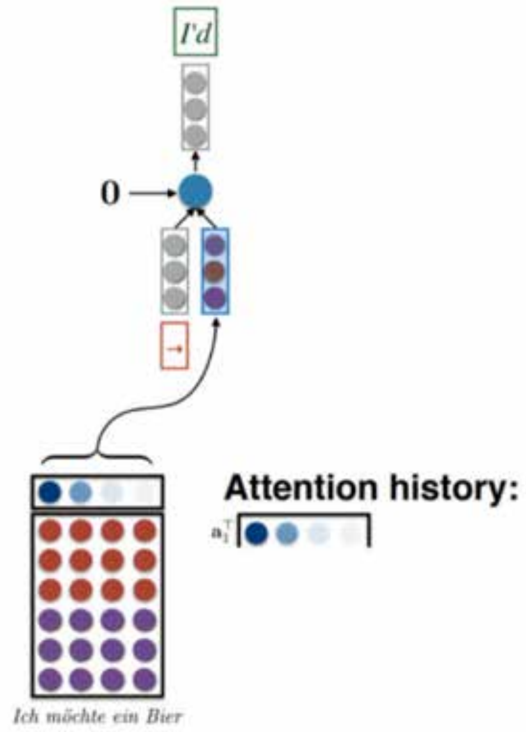
Decoding: RNN + Attention



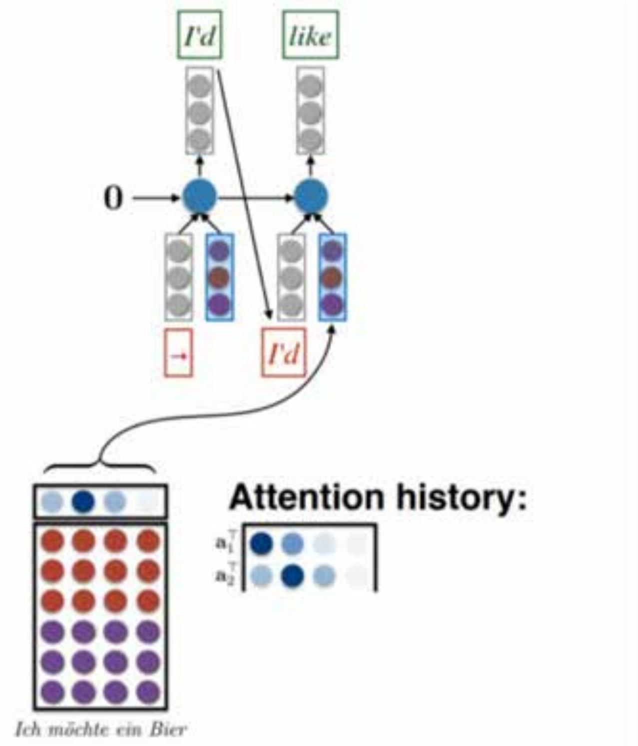
Slide Credits: Diyi Yang (Georgia Tech)



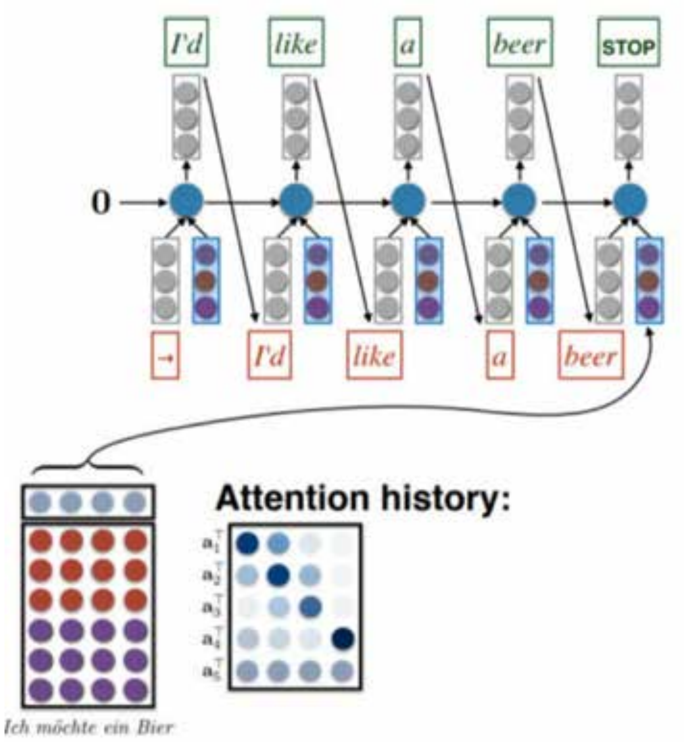
Slide Credits: Diyi Yang (Georgia Tech)



Slide Credits: Diyi Yang (Georgia Tech)



Slide Credits: Diyi Yang (Georgia Tech)



Slide Credits: Diyi Yang (Georgia Tech)

Discussion on Attention

Attention significantly improves performance (in many applications)

- Allows the decoder to focus on certain parts of the source

Attention solves the bottleneck problem

- Allows the decoder to look into the source, bypassing bottleneck

Attention provides some interpretability

- By inspecting attention distribution, we can see what the decoder was focusing on

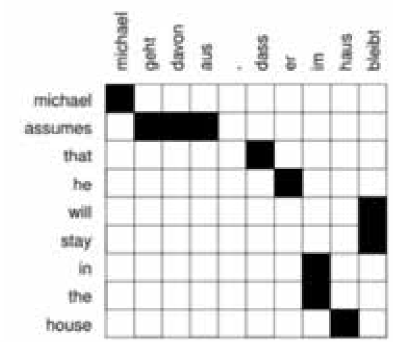
Slide Credits: Diyi Yang (Georgia Tech)

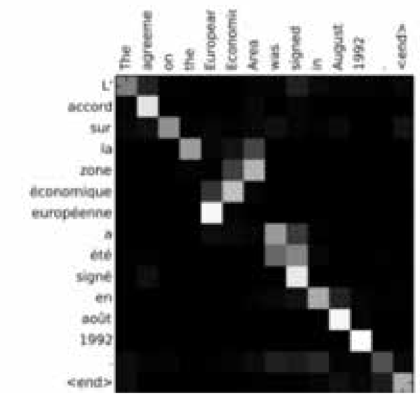
Attention vs. Alignment

Attention is similar to alignment, but there are important differences

- Alignment makes stochastic but hard decisions
  - the model picks one word or phrase at a time
- Attention is "soft" (you add together all the words)

There is no guarantee that attention corresponds to alignment since information can also flow along recurrent connections





Slide Credits: Diyi Yang (Georgia Tech)

Evaluating MT

and vs. Summarization

Slide Credits: Diyi Yang (Georgia Tech)

MT Evaluation Metrics

Manual evaluation is most accurate, but expensive

Automated evaluation metrics:

- Compare system hypothesis with reference translations
- BiLingual Evaluation Understudy (BLEU) (Papineni et al., 2002):
- Modified n-gram precision

$$p_n = \frac{\text{number of } n\text{-grams appearing in both reference and hypothesis translations}}{\text{number of } n\text{-grams appearing in the hypothesis translation}}$$

Slide Credits: Diyi Yang (Georgia Tech)

$$\text{BLEU} = \exp \frac{1}{N} \sum_{n=1}^{N} \log p_n$$

Two modifications:

- To avoid log 0, all precisions are smoothed
- Each n-gram in reference can be used at most once
  Hypothesis: *to to to to to* vs Reference: *to be or not to be*

should not get a unigram precision of 1

Precision-based metrics favor short translations

Solution: Multiply score with a brevity penalty (BP) for translations shorter than reference, $e^{1-r/h}$

Slide Credits: Diyi Yang (Georgia Tech)

BLEU Example

	Translation	p_1	p_2	p_3	p_4	BP	BLEU
Reference	*Vinay likes programming in Python*						
Sys1	*To Vinay it like to program Python*	$\frac{2}{7}$	0	0	0	1	.21
Sys2	*Vinay likes Python*	$\frac{3}{3}$	$\frac{1}{2}$	0	0	.51	.33
Sys3	*Vinay likes programming in his pajamas*	$\frac{4}{6}$	$\frac{3}{5}$	$\frac{2}{4}$	$\frac{1}{3}$	1	.76

Slide Credits: Diyi Yang (Georgia Tech)

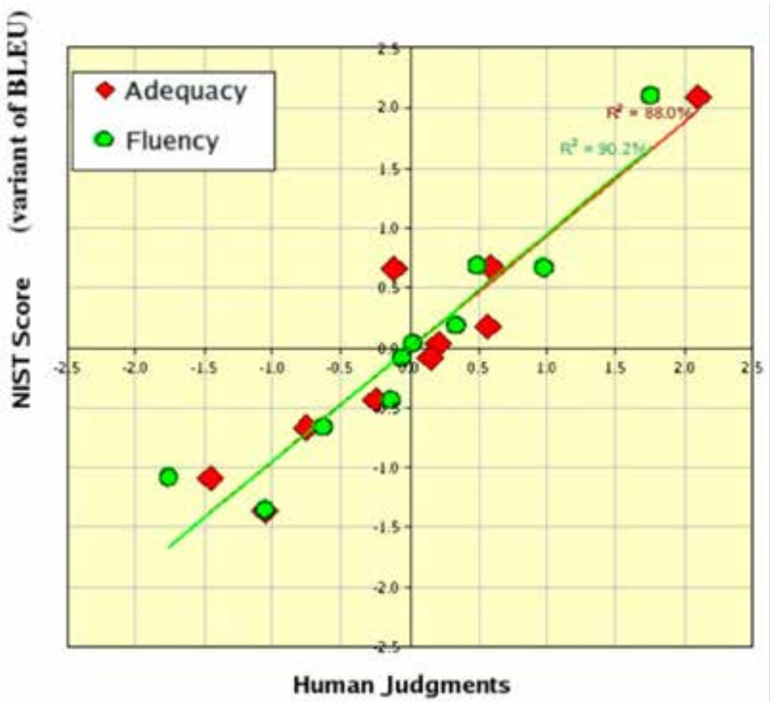
… Vs. ROUGE?

ROUGE for summarization is a complementary evaluation metric.

It measures n-gram recall from the reference summaries.

Both metrics correlate with humans

… somewhat well



Alternatives have been proposed:

- MT: METEOR: weighted F-measure
- MT: Translation Error Rate (TER): Edit distance between hypothesis and reference
- Summarization: Pyramid: hierarchical nugget recall.

Slide Credits: Diyi Yang (Georgia Tech)

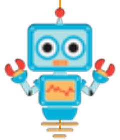
Question Answering II

Direct Modeling

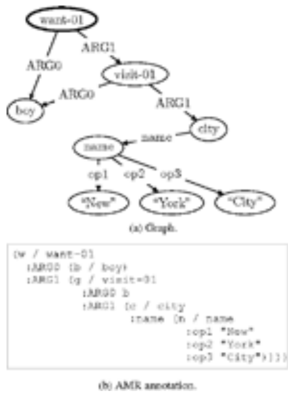
Symbolic Approaches (until ~2014)

In January 1880, two of Tesla's uncles put together enough money to help him leave Gospić for Prague where he was to study. Unfortunately, he arrived too late to enrol at Charles-Ferdinand University; he never studied Greek, a required subject; and he was illiterate in Czech, another required subject. Tesla did, however, attend lectures at the university, although, as an auditor, he did not receive grades for the courses.

[Passage of Text]



converts into



[Meaning]

uses for

[Information Need]

Feature Based Methods

Generate a list of candidate answers $A = (a_1, a_2, \ldots, a_M)$

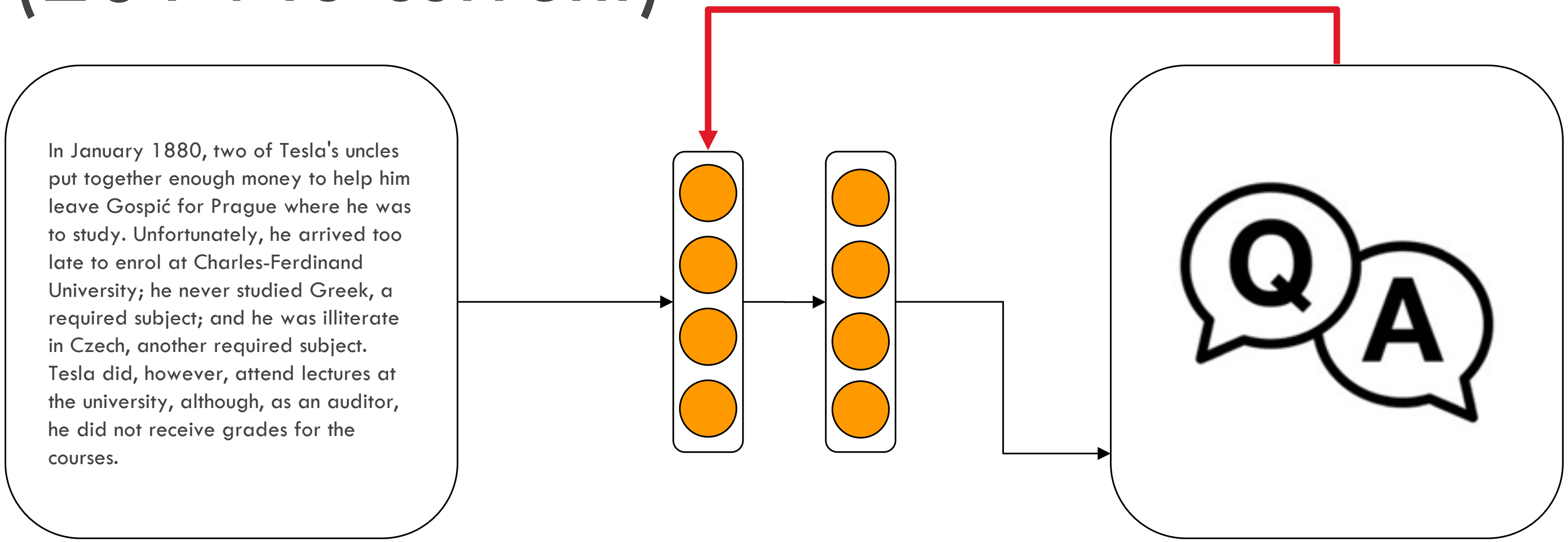
Define a feature vector $\phi(passage, question, candidate) \in \mathbb{R}^d$

- Word/bigram features
- Parse tree matches
- Dependency labels, length, part-of-speech tags

Apply a multi-class logistic regression model

Slide Credits: Diyi Yang (Georgia Tech)

End-to-End Approaches (2014 to current)



[Passage of Text] [Meaning?] [Information Need]

Creating large scale training data

Via entity anonymization

Original Version	Anonymised Version
Context	
The BBC producer allegedly struck by Jeremy Clarkson will not press charges against the "Top Gear" host, his lawyer said Friday. Clarkson, who hosted one of the most-watched television shows in the world, was dropped by the BBC Wednesday after an internal investigation by the British broadcaster found he had subjected producer Oisin Tymon "to an unprovoked physical and verbal attack." . . .	the *ent381* producer allegedly struck by *ent212* will not press charges against the " *ent153* " host , his lawyer said friday . *ent212* , who hosted one of the most - watched television shows in the world , was dropped by the *ent381* wednesday after an internal investigation by the *ent180* broadcaster found he had subjected producer *ent193* " to an unprovoked physical and verbal attack . " . . .
Query	
Producer **X** will not press charges against Jeremy Clarkson, his lawyer says.	producer **X** will not press charges against *ent212* , his lawyer says .
Answer	
Oisin Tymon	*ent193*

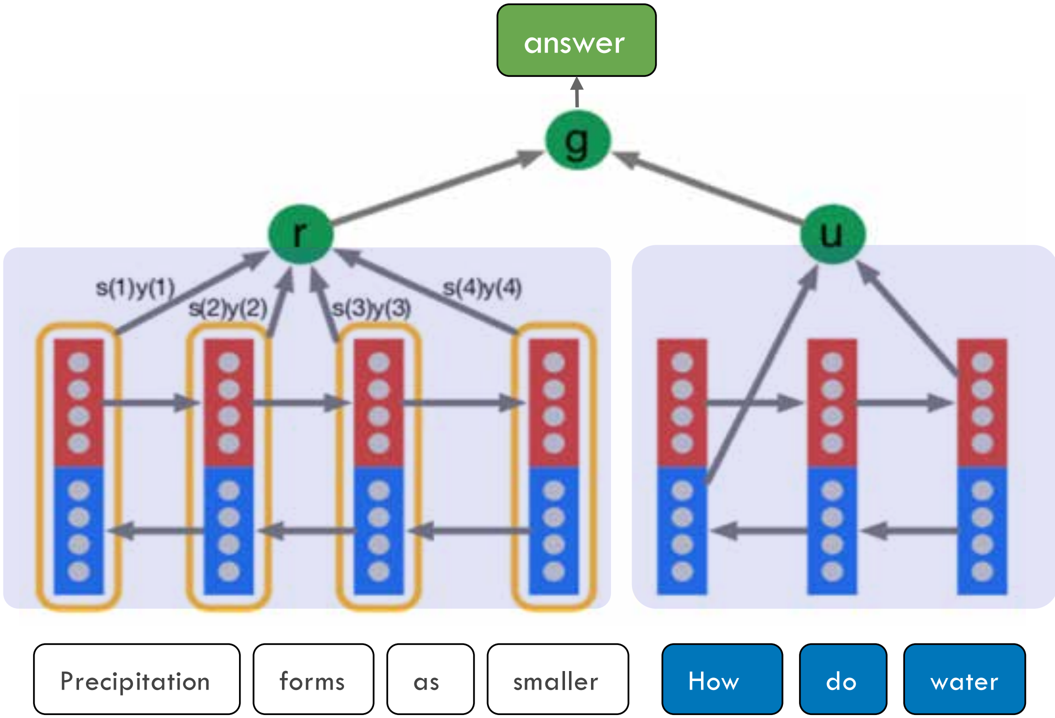
Table 3: Original and anonymised version of a data point from the Daily Mail validation set. The anonymised entity markers are constantly permuted during training and testing.

The Attentive Reader Model: Overview

Early neural model for Machine Reading

Main components reused in many other models

Hermann *et al.* (2015), *Teaching Machines to Read and Comprehend*



Slide Credits: Diyi Yang (Georgia Tech)

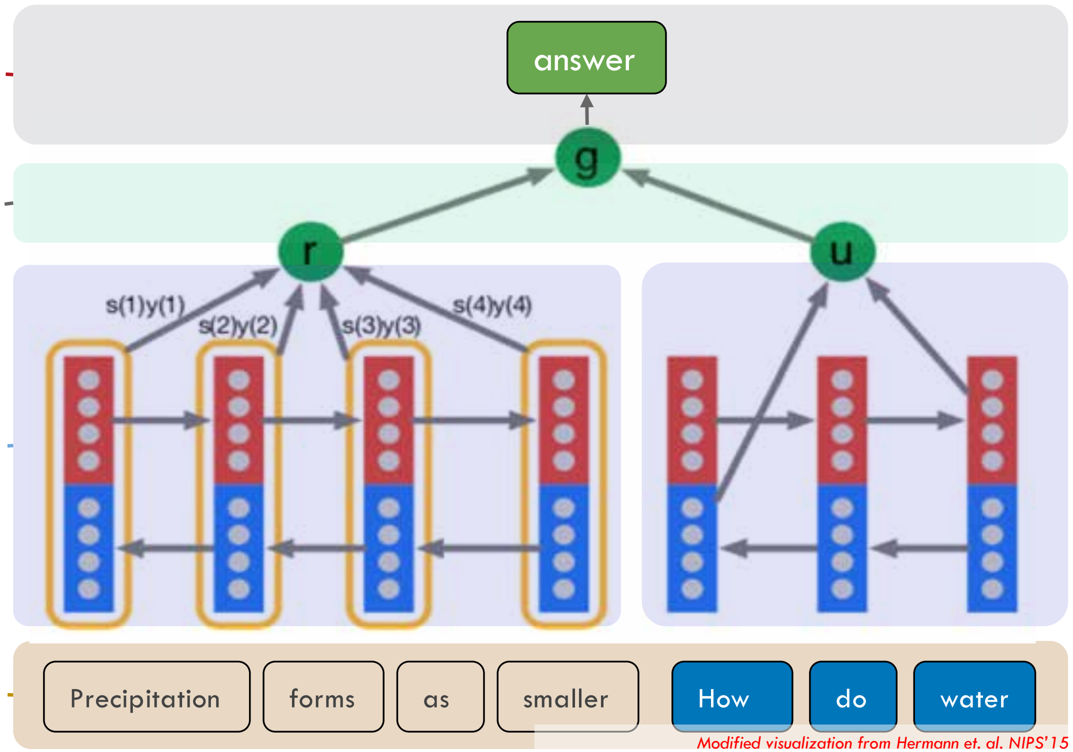
The Attentive Reader Model: Overview

Answer Selection:
answer prediction

Sequence Interaction:
Matching text with
question

Composition:
incorporating context
around words

Input: Representing
symbols as vectors



The Attentive Reader

Denote the outputs of a bidirectional LSTM as $\overrightarrow{y}(t)$ and $\overleftarrow{y}(t)$. Form two encodings, one for the query and one for each token in the document

$$u = \overrightarrow{y}_q(|q|)||\ \overleftarrow{y}_q(1) \qquad\qquad y_d(t) = \overrightarrow{y}_d(t)||\ \overleftarrow{y}_d(t)$$

The representation r of the document d is formed by a weighted sum of the token vectors. The weights are interpreted as the model's attention.

$$r = y_d \cdot s$$

$$s(t) \propto \exp\big(W_{ms} m(t)\big)$$

$$m(t) = \tanh\big(W_{ym} y_d(t) + W_{um} u\big)$$

Define the joint document and query embedding via a non-linear combination:

$$g^{AR}(d, q) = \tanh(W_{rg} r + W_{ug} u)$$

Slide Credits: Diyi Yang (Georgia Tech)

QA as Span Selection

SQuAD (Span Selection)

The first recorded travels by Europeans to China and back date from this time. The most famous traveler of the period was the Venetian Marco Polo, whose account of his trip to "Cambaluc," the capital of the Great Khan, and of life there astounded the people of Europe. The account of his travels, Il milione (or, The Million, known in English as the Travels of Marco Polo), appeared about the year 1299. Some argue over the accuracy of Marco Polo's accounts due to the lack of mentioning the Great Wall of China, tea houses, which would have been a prominent sight since Europeans had yet to adopt a tea culture, as well the practice of foot binding by the women in capital of the Great Khan. Some suggest that Marco Polo acquired much of his knowledge through contact with Persian traders since many of the places he named were in Persian.

How did some suspect that Polo learned about China instead of by actually visiting it?

Answer: through contact with Persian traders

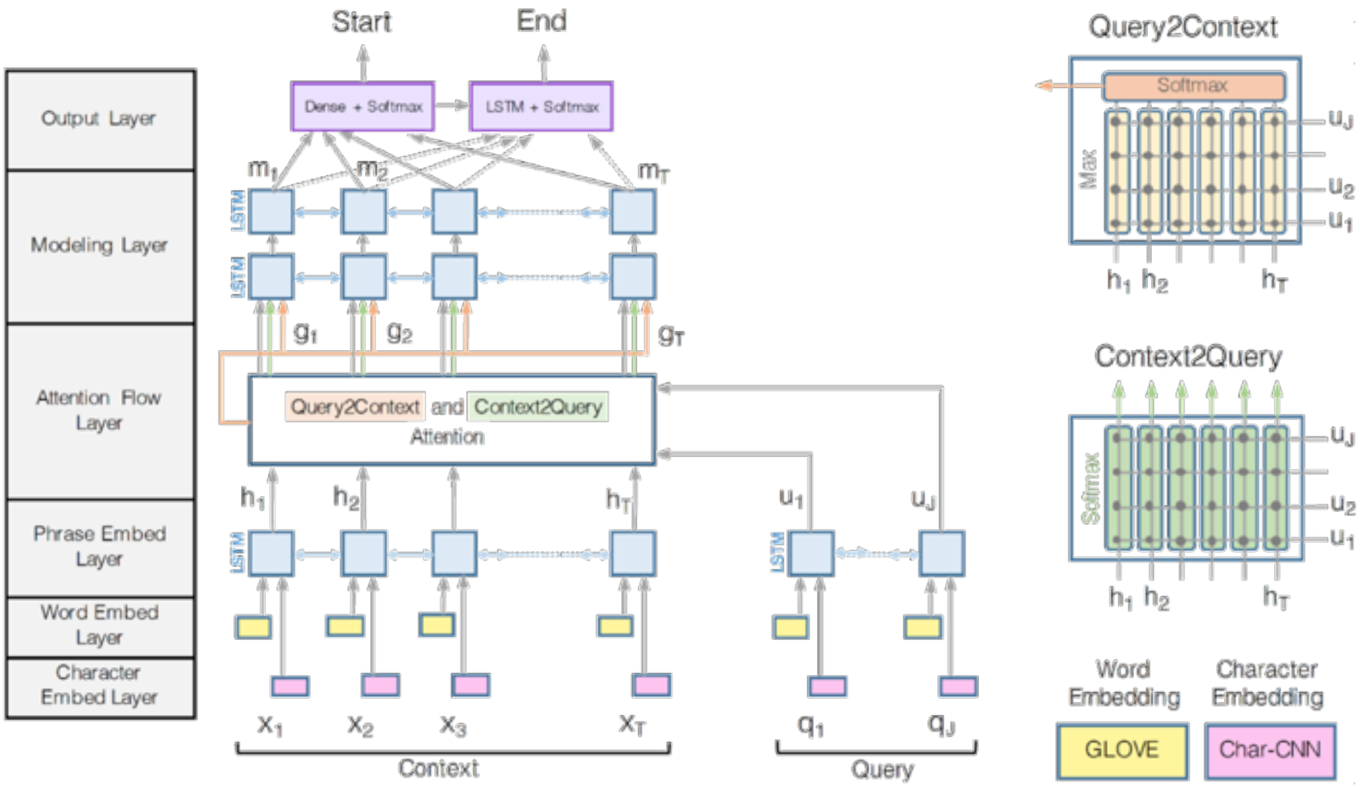
- (passage, question, answer) triples
- Passage is from Wikipedia, question is crowd-sourced
- Answer must be a span of text in the passage (aka. “extractive question answering”)
- SQuAD 1.1: 100k answerable questions, SQuAD 2.0: another 50k unanswerable questions

Rajpurkar et al. (2016) *SQuAD: 100,000+ Questions for Machine Comprehension of Text*

Slide Credits: Diyi Yang (GeorgiaTech)

BiDAF

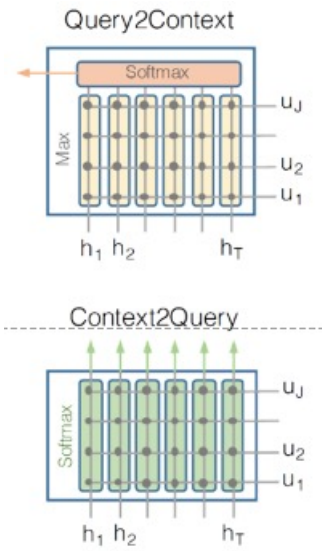
(Seo et al., 2017): *Bidirectional Attention Flow for Machine Comprehension*



Slide Credits: Diyi Yang (Georgia Tech)

BiLSTM-based Models (i.e., BIDAF)

- Encode the question using word/char embeddings; pass onto a biLSTM encoder
- Encode the passage similarly
- Passage-to-question and question-to-passage attention
- Modeling layer: another BiLSTM layer
- Output layer: two classifiers for predicting start and end points
- The entire model can be trained in an end-to-end way



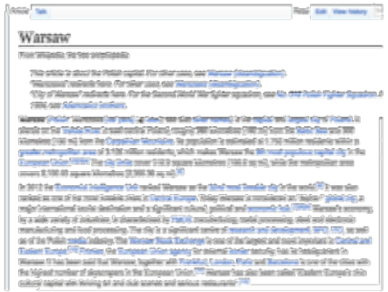
Slide Credits: Diyi Yang (Georgia Tech)

Open-domain QA

SQuAD, TREC, WebQuestions, WikiMovies

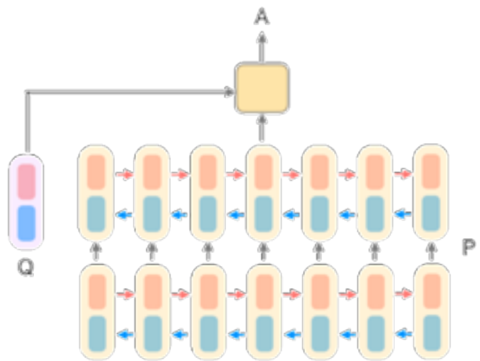
Q: How many of Warsaw's inhabitants spoke Polish in 1933?

Document Retriever



Document Reader

833,500



Chen *et al.* (2017) *Reading Wikipedia to answer open-domain questions*

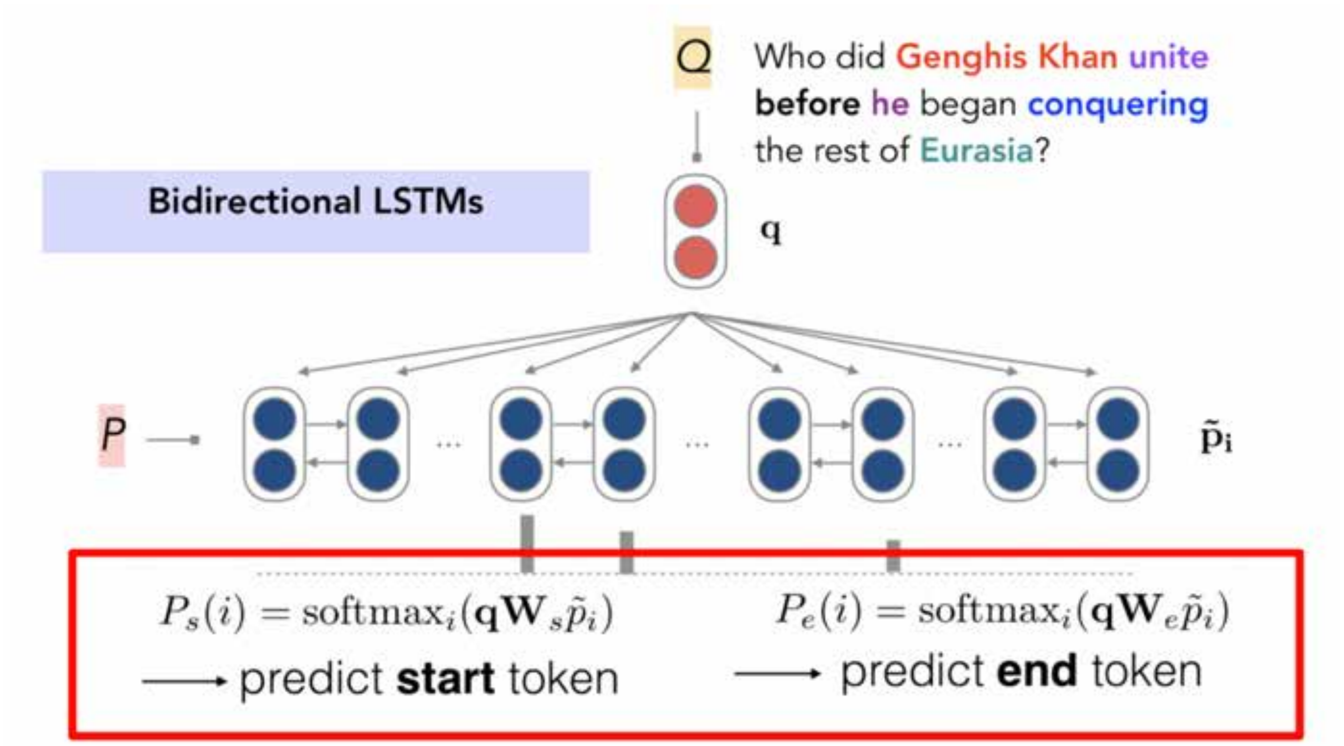
Document Retriever: Two Steps

1. $tf.idf$ bag-of-words vector representation

2. Efficient bigram hashing

Slide Credits: Diyi Yang (Georgia Tech)

Document Reader



Slide Credits: Diyi Yang (Georgia Tech)

Document Reader: Prediction

Goal: predict the span of tokens that is most likely the correct answer:

$$\max_{i,j} P_{start}(i) \times P_{end}(j)$$

Train two classifiers independently for predicting ends of span

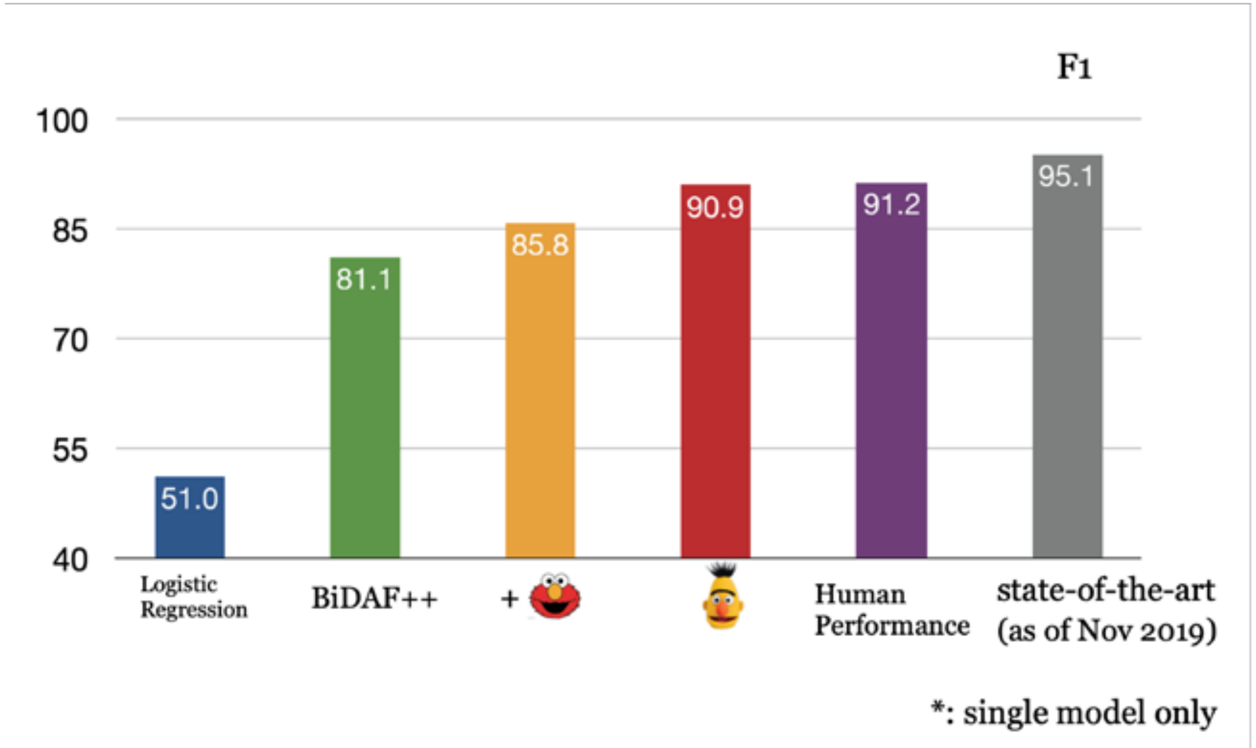
but constrained such that $i \leq j \leq i + 15$ and

$P_{start}(i), P_{end}(j)$ is the probability of each token being start, end.

- $P_{end}(i) \propto \exp(p_i W_s q)$
- $P_{end}(i) \propto \exp(p_i W_e q)$

Slide Credits: Diyi Yang (Georgia Tech)

SOLVED! … or not?



Slide Credits: Diyi Yang (Georgia Tech)

Is Reading Comprehension Solved?

Article: Super Bowl 50
Paragraph: "*Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV.*"
Question: "*What is the name of the quarterback who was 38 in Super Bowl XXXIII?*"
Original Prediction: John Elway
Prediction under adversary: Jeff Dean

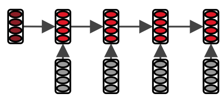
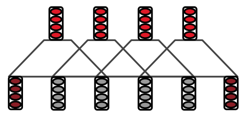
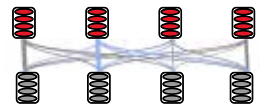
Models are brittle: *Easy to create adversarial examples*

(Jia et al, 2017): *Adversarial Examples for Evaluating Reading Comprehension Systems*

Slide Credits: Diyi Yang (Georgia Tech)

Compositional Sequence Encoders Overview

Language is compositional! Characters → Words → Phrases → Clauses → Sentences → Paragraphs → Documents

Architecture	RNN (LSTM, GRU)	CNN	Self-Attention
Illustration			
Function $\mathbf{y}_t =$	$f(\mathbf{x}_t, \mathbf{y}_{t-1})$	$f(\mathbf{x}_{t-k}, \ldots, \mathbf{x}_{t+k})$	$f(\mathbf{x}_1, \ldots, \mathbf{x}_T)$
Advantages	- unlimited context - recency bias	- parallelizable → fast - local n-gram patterns	- parallelizable → fast - long-range dependencies

More to learn: Transformers