Sequence to Sequence (seq2seq) & Attention

Ling 282/482: Deep Learning for Computational Linguistics

C.M. Downey

Fall 2025

seq2seq: Overview

Sequence-to-sequence problems

Sequence-to-sequence problems

- seq2seq: input a sequence and output a **different sequence**

Sequence-to-sequence problems

- seq2seq: input a sequence and output a **different sequence**
- Many NLP tasks can be framed as **sequence-to-sequence** problems

Sequence-to-sequence problems

- seq2seq: input a sequence and output a **different sequence**
- Many NLP tasks can be framed as **sequence-to-sequence** problems
  - **Machine Translation**: sequence of source language tokens to sequence of target language tokens

Sequence-to-sequence problems

- seq2seq: input a sequence and output a **different sequence**
- Many NLP tasks can be framed as **sequence-to-sequence** problems
  - **Machine Translation**: sequence of source language tokens to sequence of target language tokens
  - **Parsing**: “Shane talks.” —> “(S (NP (N Shane)) (VP V talks))”

Sequence-to-sequence problems

- seq2seq: input a sequence and output a **different sequence**
- Many NLP tasks can be framed as **sequence-to-sequence** problems
  - **Machine Translation**: sequence of source language tokens to sequence of target language tokens
  - **Parsing**: “Shane talks.” —> “(S (NP (N Shane)) (VP V talks))”
    - Semantic as well as syntactic

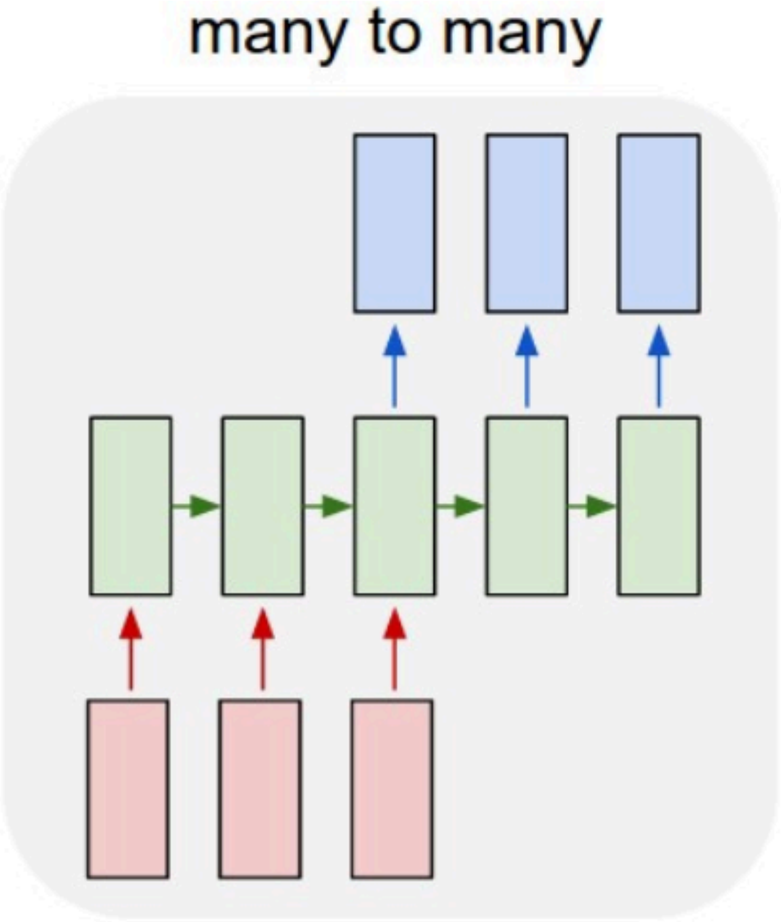
Sequence-to-sequence problems

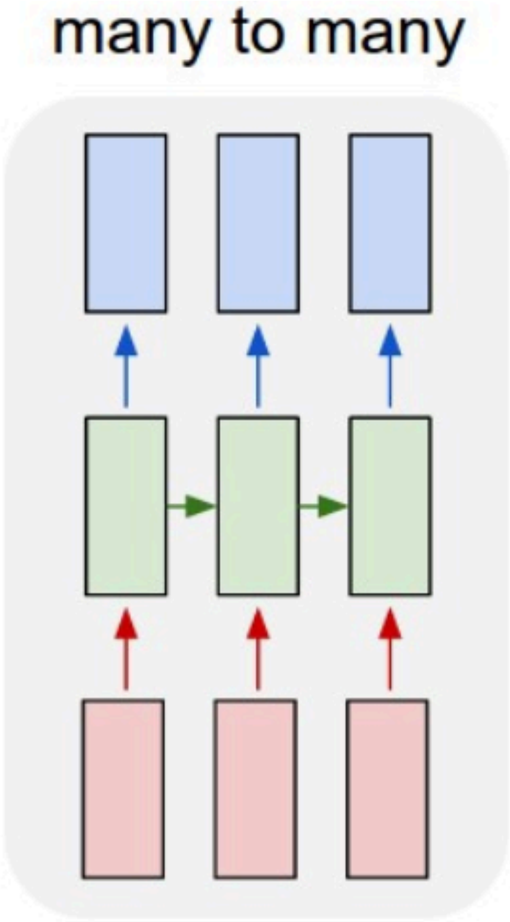
- seq2seq: input a sequence and output a **different sequence**
- Many NLP tasks can be framed as **sequence-to-sequence** problems
  - **Machine Translation**: sequence of source language tokens to sequence of target language tokens
  - **Parsing**: “Shane talks.” —> “(S (NP (N Shane)) (VP V talks))”
    - Semantic as well as syntactic
  - **Summarization** and **Question Answering**

Sequence-to-sequence problems

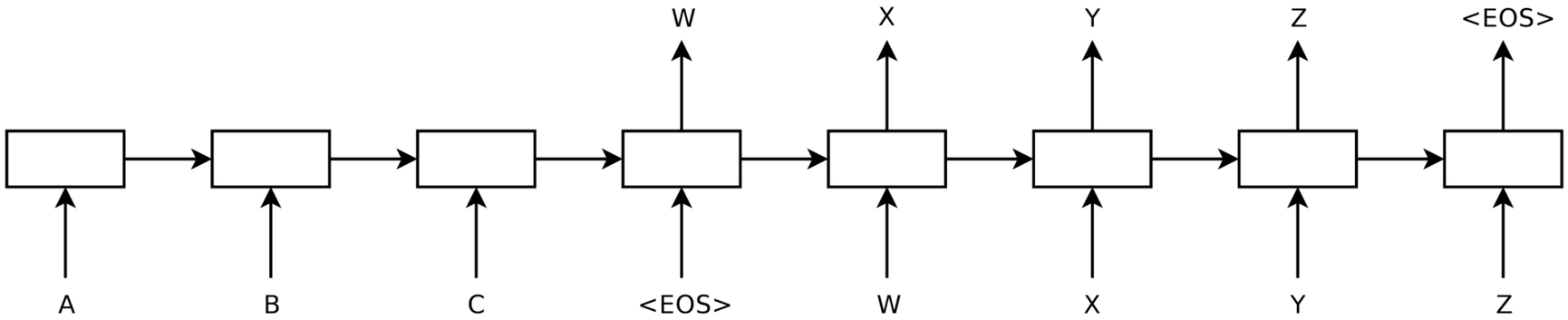
- seq2seq: input a sequence and output a **different sequence**
- Many NLP tasks can be framed as **sequence-to-sequence** problems
  - **Machine Translation**: sequence of source language tokens to sequence of target language tokens
  - **Parsing**: “Shane talks.” —> “(S (NP (N Shane)) (VP V talks))”
    - Semantic as well as syntactic
  - **Summarization** and **Question Answering**
- **Not the same** as *tagging*, which assigns a label to each position in a given sequence

Seq2seq vs Tagging



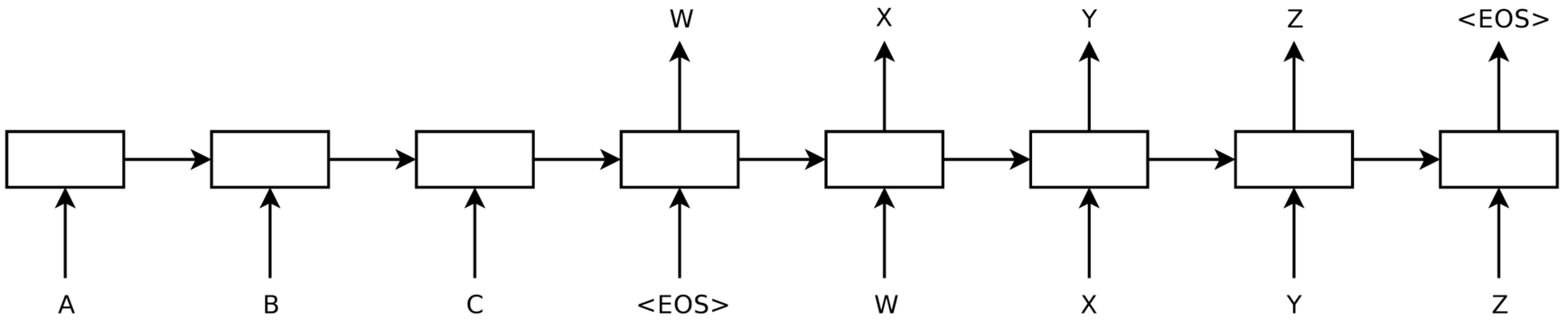


seq2seq architecture



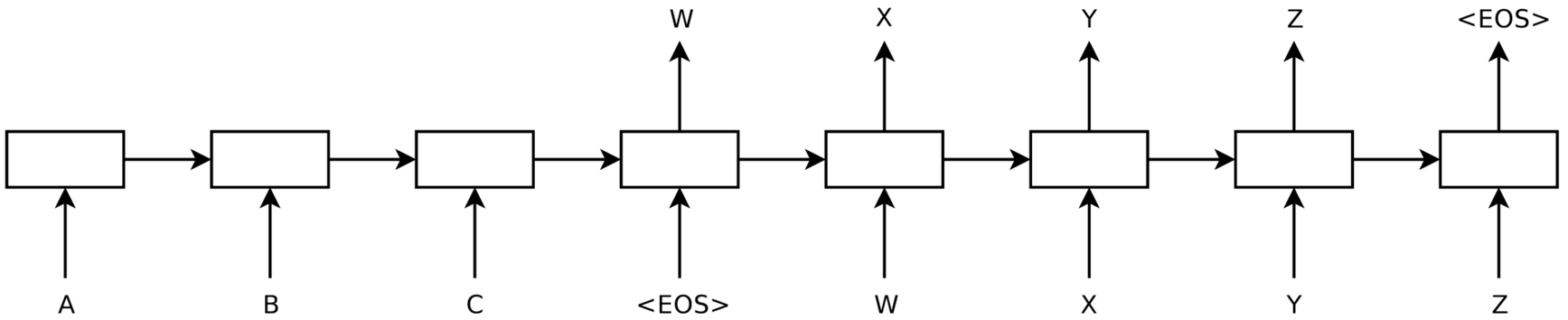
Sutskever et al 2013

seq2seq architecture



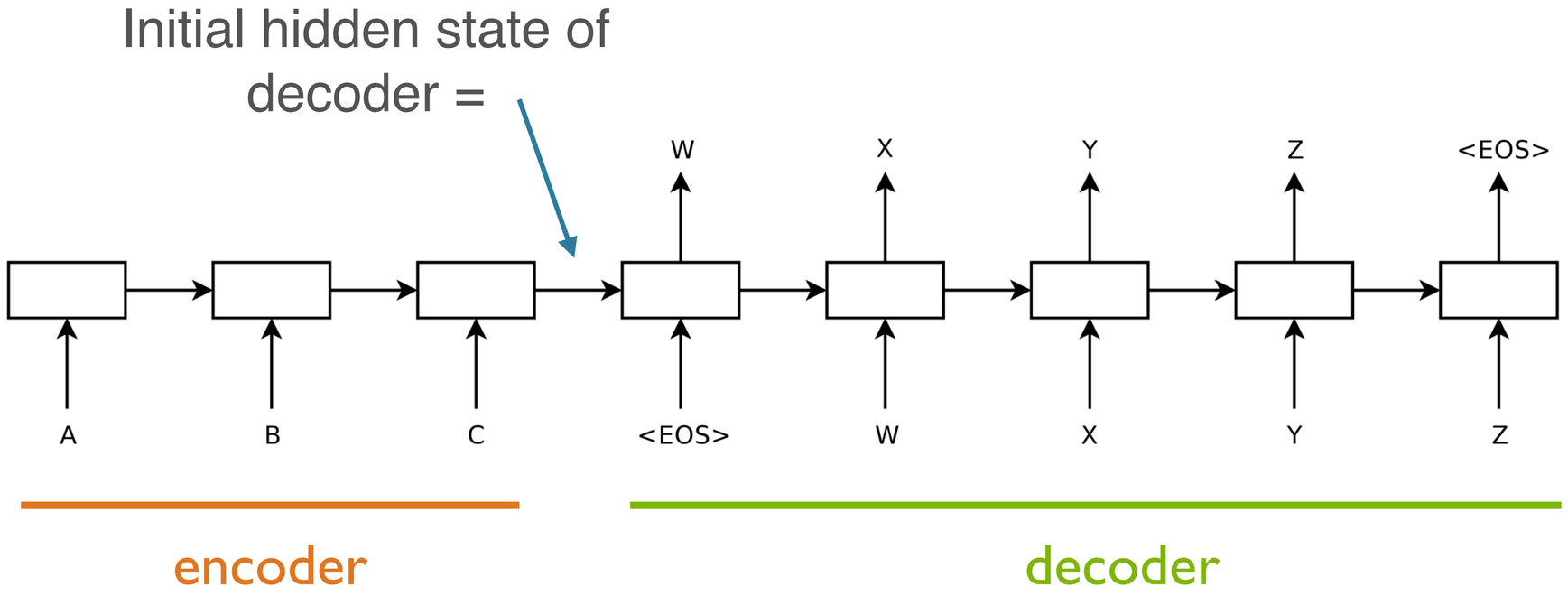
Sutskever et al 2013

seq2seq architecture



Sutskever et al 2013

seq2seq architecture



Sutskever et al 2013

seq2seq architecture

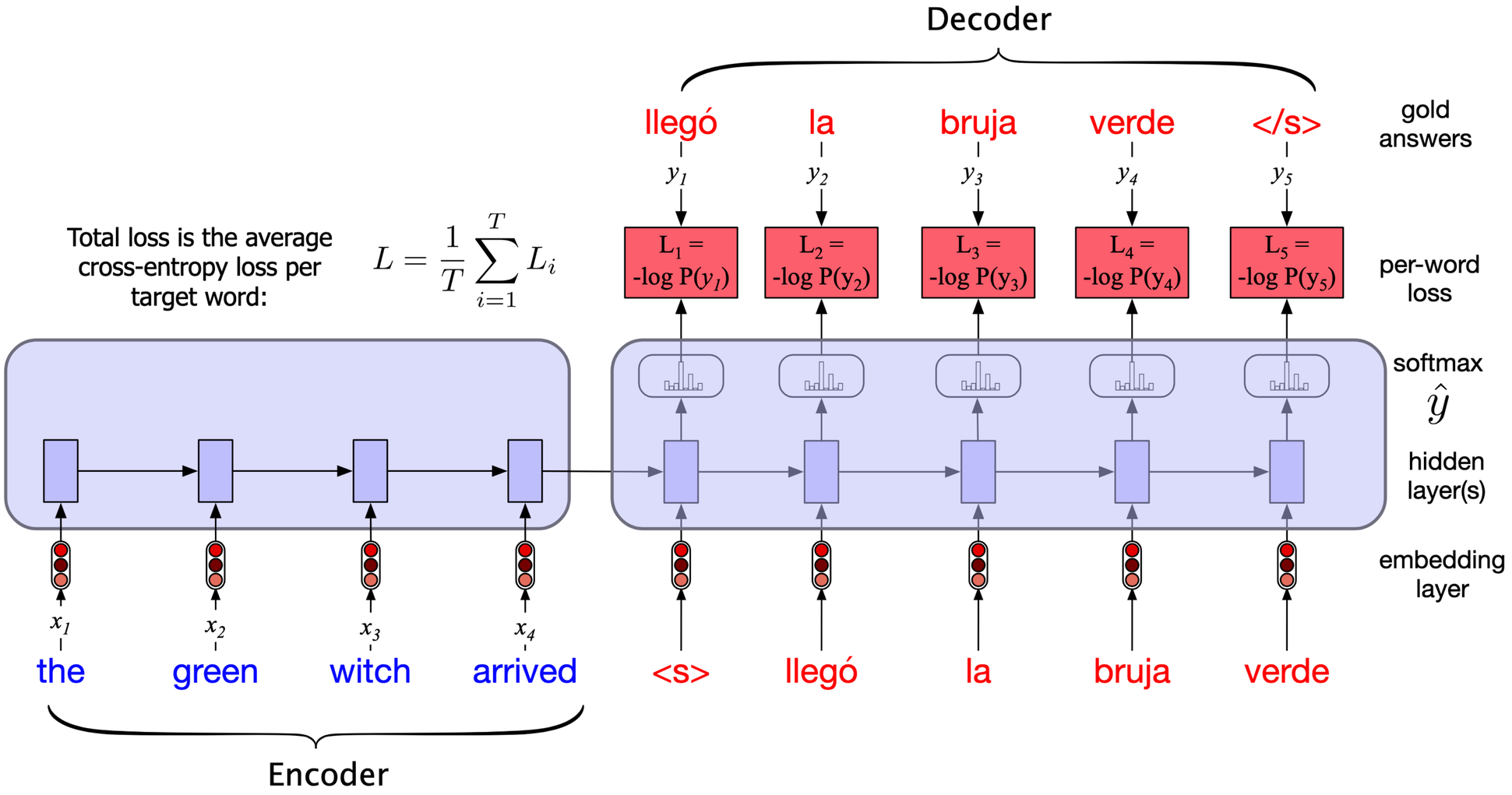
seq2seq architecture

- Two components
  - **Encoder**
    - Input sequence —> vector representation (“context” vector)
  - **Decoder**
    - Vector (“context” vector) —> Output sequence

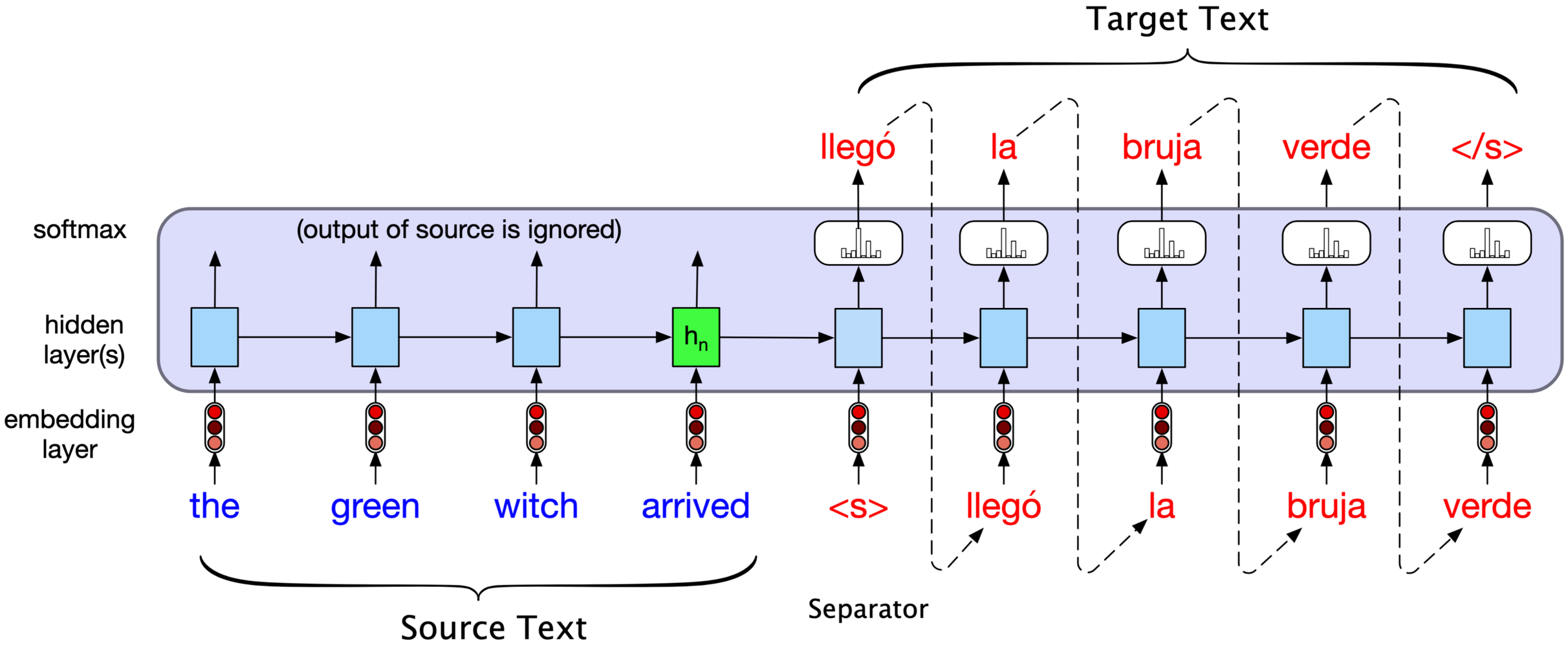
seq2seq architecture

- Two components
  - **Encoder**
    - Input sequence —> vector representation (“context” vector)
  - **Decoder**
    - Vector (“context” vector) —> Output sequence
- High-level “**API**”
  - Encoder/decoder can be **different architectures** (LSTM, GRU, Transformer, convolutional, …)

Training an encoder-decoder RNN

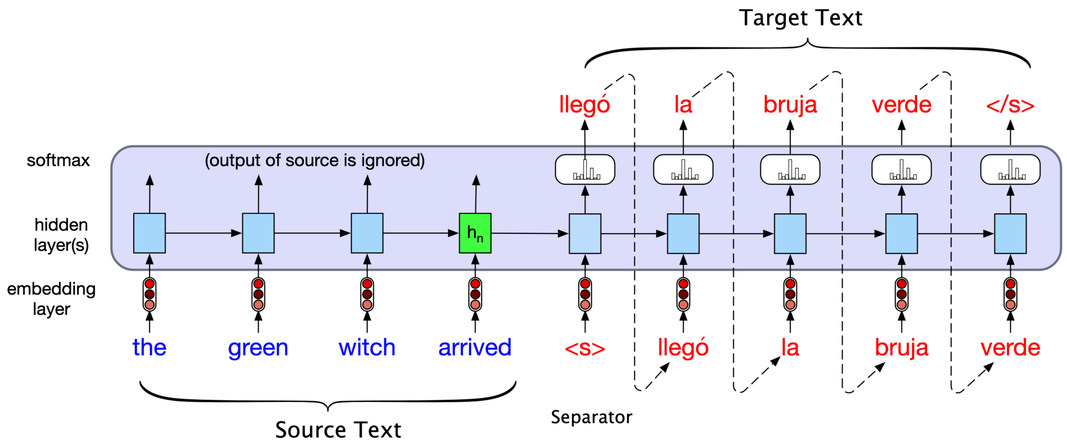


Inference / Generation



Conditional LM

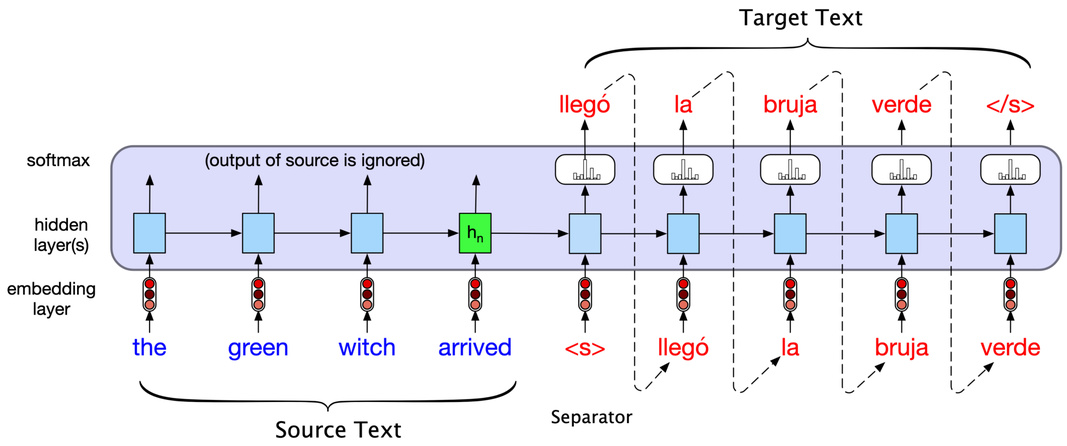
$$P(y \mid x) = \prod_{i=1}^{|y|} P(y_i \mid x, y_{<i})$$



Conditional LM

- Effectively, a seq2seq model is a **Conditional Language Model**

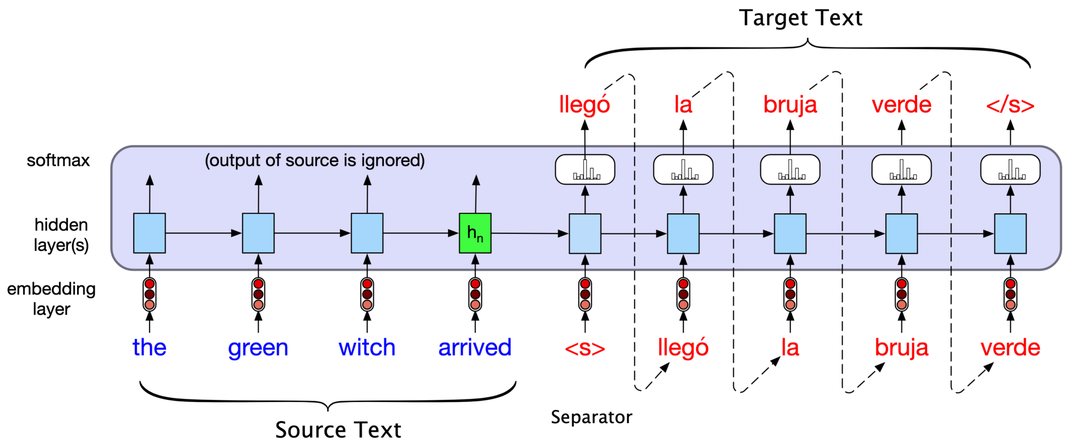
$$P(y \mid x) = \prod_{i=1}^{|y|} P(y_i \mid x, y_{<i})$$



Conditional LM

- Effectively, a seq2seq model is a **Conditional Language Model**
- LMs like we have seen, but **conditioned on the input**

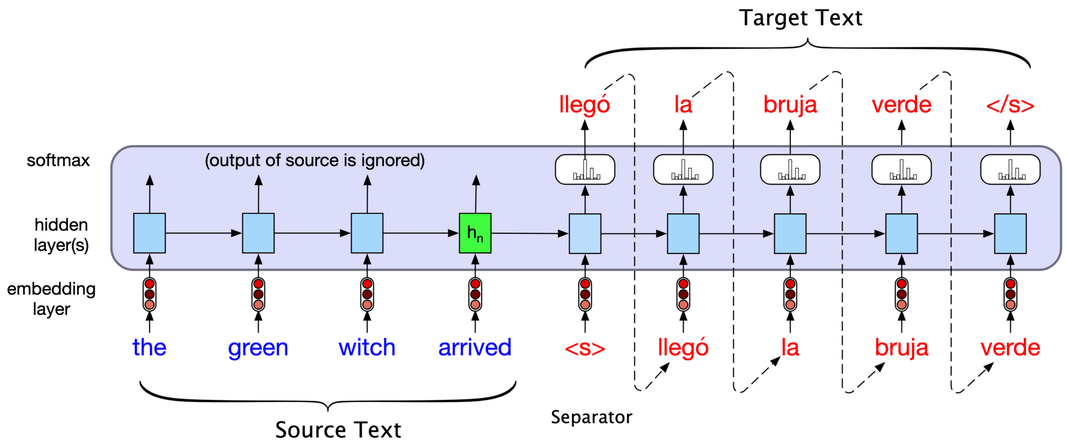
$$P(y \mid x) = \prod_{i=1}^{|y|} P(y_i \mid x, y_{<i})$$



Conditional LM

- Effectively, a seq2seq model is a **Conditional Language Model**
- LMs like we have seen, but **conditioned on the input**
- LMs were already conditioned on the **output sequence prefix**

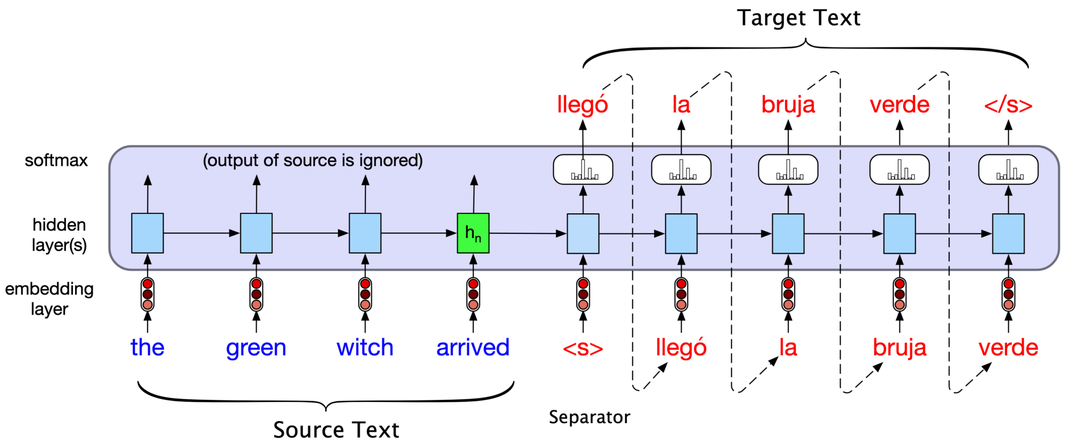
$$P(y \mid x) = \prod_{i=1}^{|y|} P(y_i \mid x, y_{<i})$$



Conditional LM

- Effectively, a seq2seq model is a **Conditional Language Model**
- LMs like we have seen, but **conditioned on the input**
- LMs were already conditioned on the **output sequence prefix**
- Each step of the output is conditioned on the **whole** of the input

$$P(y \mid x) = \prod_{i=1}^{|y|} P(y_i \mid x, y_{<i})$$



Translation Evaluation

Source

la verdad, cuya madre es la historia, émula del tiempo, depósito de las acciones,
testigo de lo pasado, ejemplo y aviso de lo presente, advertencia de lo por venir.

Reference

truth, whose mother is history, rival of time, storehouse of deeds,
witness for the past, example and counsel for the present, and warning for the future.

Candidate 1

truth, whose mother is history, voice of time, deposit of actions,
witness for the past, example and warning for the present, and warning for the future

Candidate 2

the truth, which mother is the history, émula of the time, deposition of the shares,
witness of the past, example and notice of the present, warning of it for coming

JM S11.8

Translation Evaluation

- **Ideal: human evaluation** (fluency, adequacy, ranking)

Source

la verdad, cuya madre es la historia, émula del tiempo, depósito de las acciones,
testigo de lo pasado, ejemplo y aviso de lo presente, advertencia de lo por venir.

Reference

truth, whose mother is history, rival of time, storehouse of deeds,
witness for the past, example and counsel for the present, and warning for the future.

Candidate 1

truth, whose mother is history, voice of time, deposit of actions,
witness for the past, example and warning for the present, and warning for the future

Candidate 2

the truth, which mother is the history, émula of the time, deposition of the shares,
witness of the past, example and notice of the present, warning of it for coming

JM S11.8

Translation Evaluation

- **Ideal: human evaluation** (fluency, adequacy, ranking)
- **BLEU** (BiLingual Evaluation Understudy): roughly, **n-gram overlap** between reference translations and machine translations

Source

la verdad, cuya madre es la historia, émula del tiempo, depósito de las acciones,
testigo de lo pasado, ejemplo y aviso de lo presente, advertencia de lo por venir.

Reference

truth, whose mother is history, rival of time, storehouse of deeds,
witness for the past, example and counsel for the present, and warning for the future.

Candidate 1

truth, whose mother is history, voice of time, deposit of actions,
witness for the past, example and warning for the present, and warning for the future

Candidate 2

the truth, which mother is the history, émula of the time, deposition of the shares,
witness of the past, example and notice of the present, warning of it for coming

JM S11.8

Translation Evaluation

- **Ideal: human evaluation** (fluency, adequacy, ranking)
- **BLEU** (BiLingual Evaluation Understudy): roughly, **n-gram overlap** between reference translations and machine translations
  - Con: **penalizes synonymous translations**

Source

la verdad, cuya madre es la historia, émula del tiempo, depósito de las acciones,
testigo de lo pasado, ejemplo y aviso de lo presente, advertencia de lo por venir.

Reference

truth, whose mother is history, rival of time, storehouse of deeds,
witness for the past, example and counsel for the present, and warning for the future.

Candidate 1

truth, whose mother is history, voice of time, deposit of actions,
witness for the past, example and warning for the present, and warning for the future

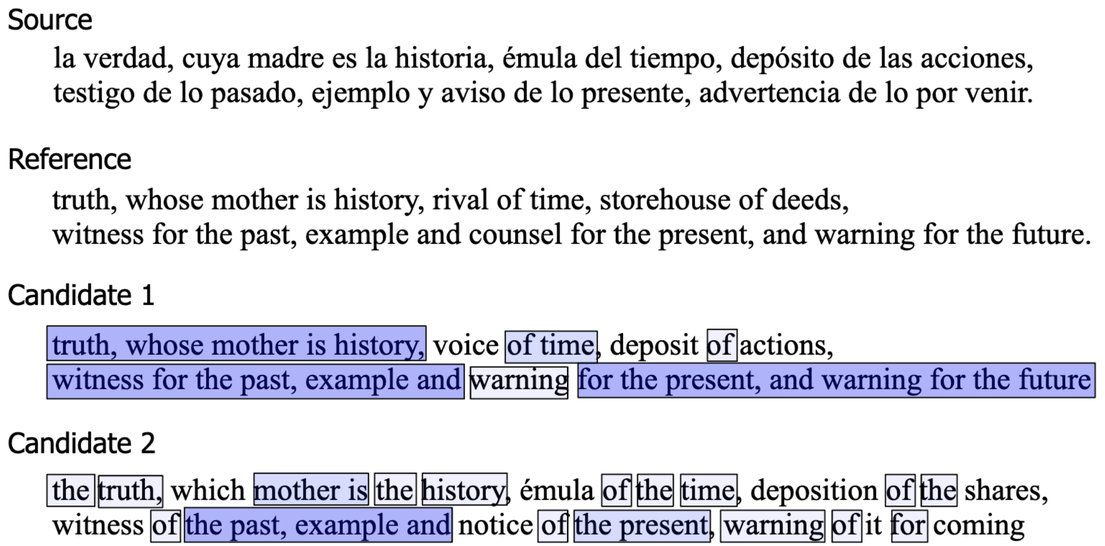
Candidate 2

the truth, which mother is the history, émula of the time, deposition of the shares,
witness of the past, example and notice of the present, warning of it for coming

JM S11.8

Translation Evaluation

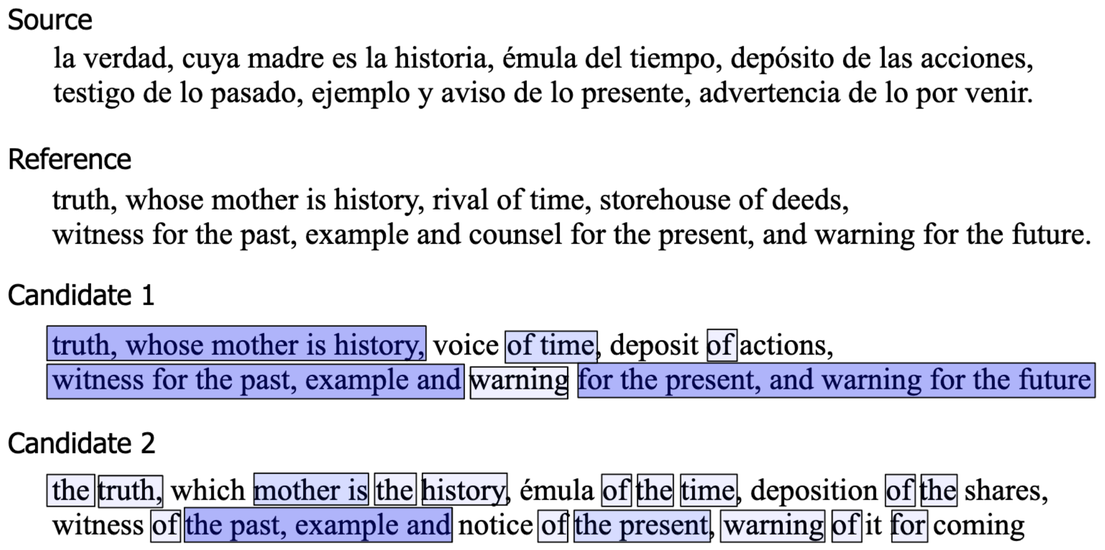
- **Ideal: human evaluation** (fluency, adequacy, ranking)
- **BLEU** (BiLingual Evaluation Understudy): roughly, **n-gram overlap** between reference translations and machine translations
  - Con: **penalizes synonymous translations**
    - METEOR, BERTScore attempt to alleviate



JM S11.8

Translation Evaluation

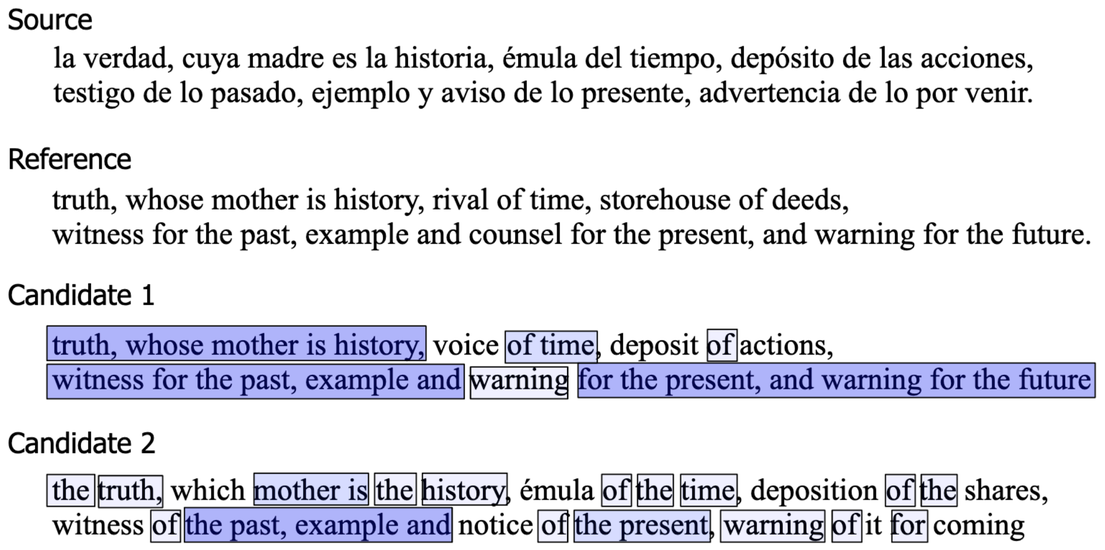
- **Ideal: human evaluation** (fluency, adequacy, ranking)
- **BLEU** (BiLingual Evaluation Understudy): roughly, **n-gram overlap** between reference translations and machine translations
  - Con: **penalizes synonymous translations**
    - METEOR, BERTScore attempt to alleviate
  - **Low correlation** with human ratings



JM S11.8

Translation Evaluation

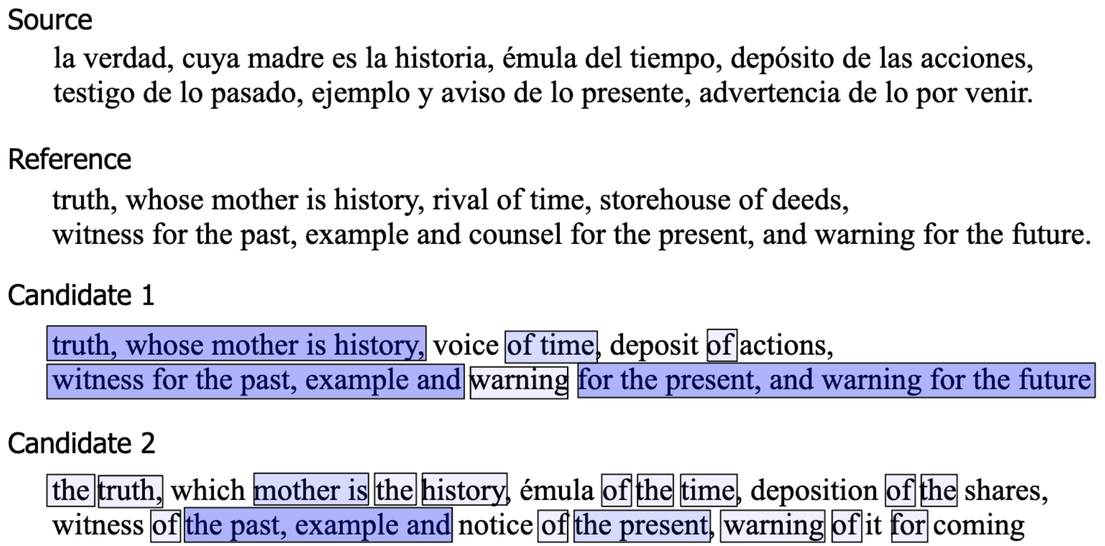
- **Ideal: human evaluation** (fluency, adequacy, ranking)
- **BLEU** (BiLingual Evaluation Understudy): roughly, **n-gram overlap** between reference translations and machine translations
  - Con: **penalizes synonymous translations**
    - METEOR, BERTScore attempt to alleviate
  - **Low correlation** with human ratings
- chrF++



JM S11.8

Translation Evaluation

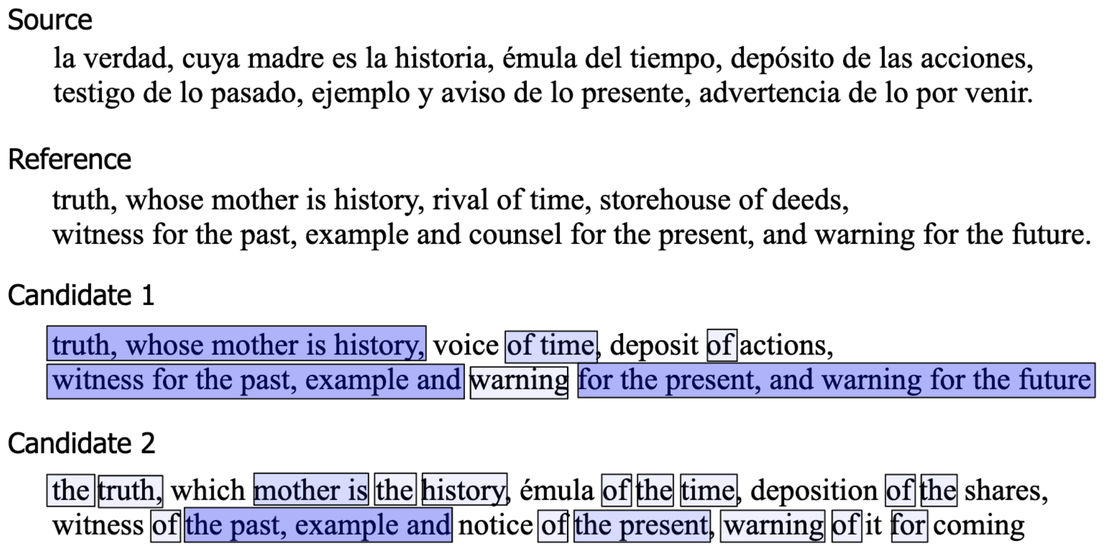
- **Ideal: human evaluation** (fluency, adequacy, ranking)
- **BLEU** (BiLingual Evaluation Understudy): roughly, **n-gram overlap** between reference translations and machine translations
  - Con: **penalizes synonymous translations**
    - METEOR, BERTScore attempt to alleviate
  - **Low correlation** with human ratings
- chrF++
  - Refinement of **character-level** n-gram F1 score



JM S11.8

Translation Evaluation

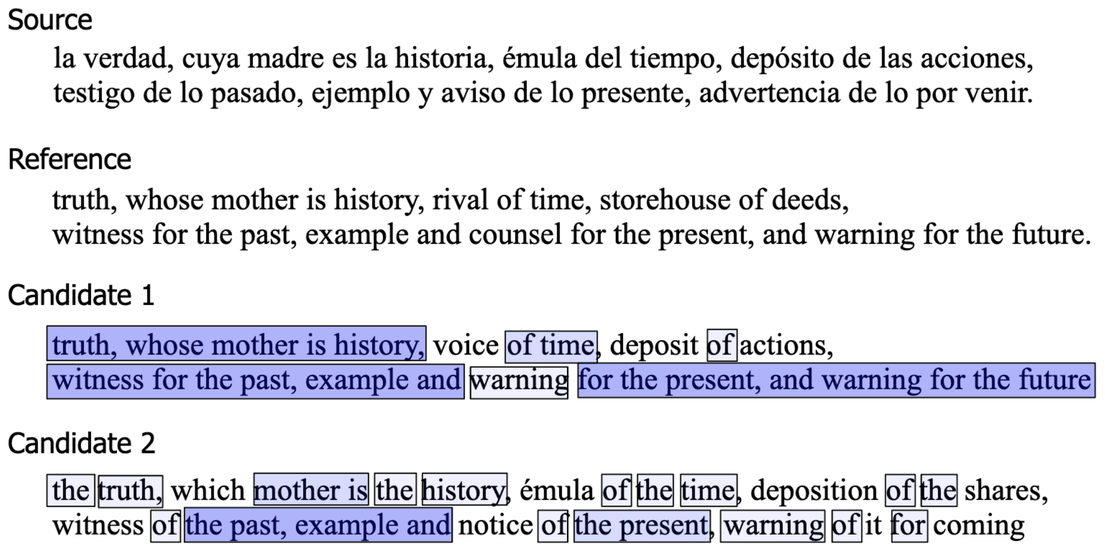
- **Ideal: human evaluation** (fluency, adequacy, ranking)
- **BLEU** (BiLingual Evaluation Understudy): roughly, **n-gram overlap** between reference translations and machine translations
  - Con: **penalizes synonymous translations**
    - METEOR, BERTScore attempt to alleviate
  - **Low correlation** with human ratings
- chrF++
  - Refinement of **character-level** n-gram F1 score
  - Seems to have better correlations



JM S11.8

Translation Evaluation

- **Ideal: human evaluation** (fluency, adequacy, ranking)
- **BLEU** (BiLingual Evaluation Understudy): roughly, **n-gram overlap** between reference translations and machine translations
  - Con: **penalizes synonymous translations**
    - METEOR, BERTScore attempt to alleviate
  - **Low correlation** with human ratings
- chrF++
  - Refinement of **character-level** n-gram F1 score
  - Seems to have better correlations
- In general: still no perfect solution



JM S11.8

Outstanding Issues in MT

- Evaluation: automated metrics are all flawed
  - "Tangled Up in BLEU"
- Low-resource / unsupervised MT
  - Can we build good translation models in the absence of huge amounts of parallel text?
    - Common technique: *backtranslation*
  - http://www.statmt.org/wmt20/unsup_and_very_low_res/
  - http://turing.iimas.unam.mx/americasnlp/st.html
  - https://www.aclweb.org/anthology/2020.acl-main.560/

Sequence Alignment

Statistical Machine Translation (90s-2010s)

Statistical Machine Translation (90s-2010s)

- Goal: find best translation *y* (e.g. English) of source sentence *x* (e.g. French)

$$\arg\max_{y} P(y \mid x)$$

Statistical Machine Translation (90s-2010s)

- Goal: find best translation *y* (e.g. English) of source sentence *x* (e.g. French)

$$\arg\max_{y} P(y \mid x)$$

- Use Bayes' Rule to decompose into two components:

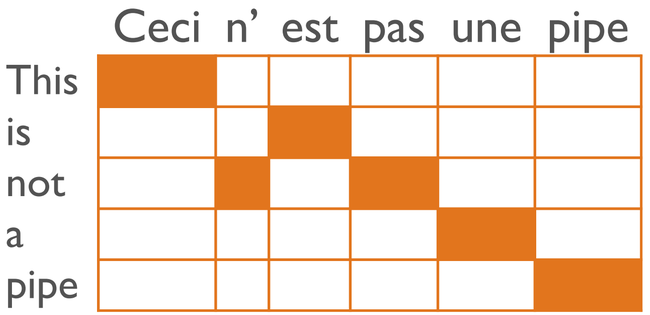
$$\arg\max_{y} P(x \mid y)P(y)$$

  - Core **translation model: P(xly)**
  - "Pure" **Language Model P(y)**: produce good / fluent target language text (e.g. English)

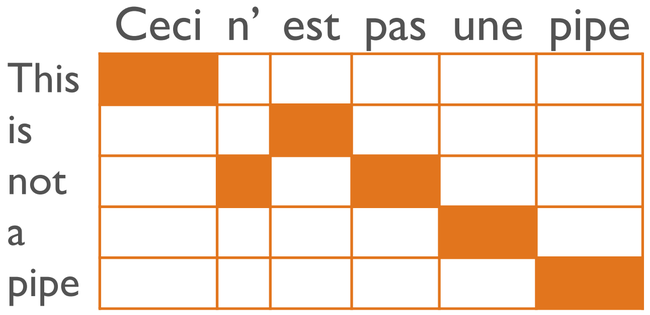
Alignment

- Most SMT systems modeled **alignment** between sequences
  - **Correspondence between words/phrases** in source and target sentence
  - Useful since languages have **very different word orders**
- Add alignment as a latent variable:

$$P(x, a \mid y)$$

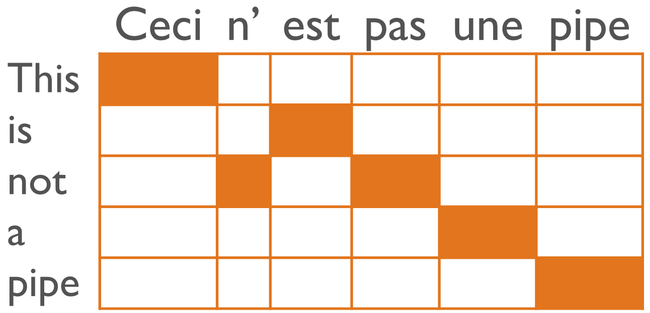


Alignment, example



Alignment, example

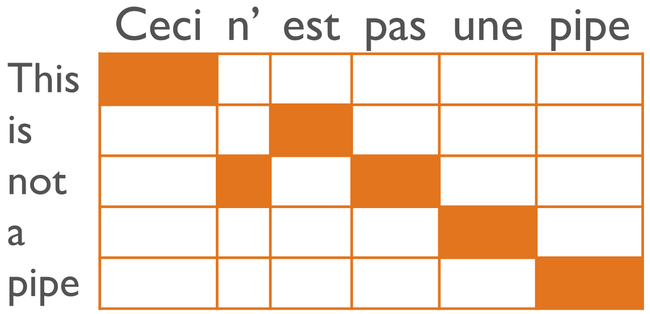
Ceci n’ est pas une pipe



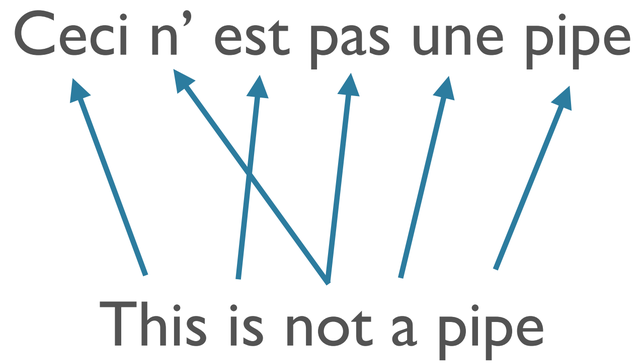
Alignment, example

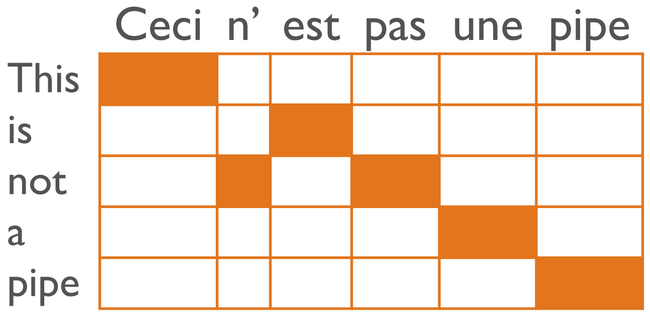
Ceci n’ est pas une pipe

This is not a pipe



Alignment, example





SMT Difficulties

SMT Difficulties

- **Key features** for determining alignment
  - Probability of **word-pairs** aligning (using a **lexicon / bilingual dictionary**)
  - Probability of a word aligning to a phrase (in general)

SMT Difficulties

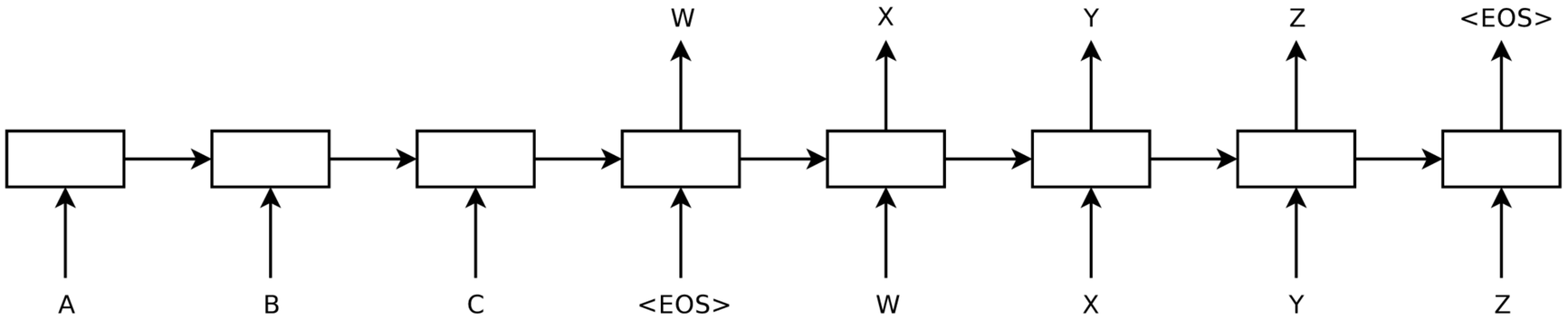
- **Key features** for determining alignment
  - Probability of **word-pairs** aligning (using a **lexicon / bilingual dictionary**)
  - Probability of a word aligning to a phrase (in general)
- Engineering hurdles:
  - Huge amounts of **hand-crafted features**
  - Reliance on **human curated resources** like dictionaries
  - Most of the above are **language-pair-specific**, have to be repeated

SMT Difficulties

- **Key features** for determining alignment
  - Probability of **word-pairs** aligning (using a **lexicon / bilingual dictionary**)
  - Probability of a word aligning to a phrase (in general)
- Engineering hurdles:
  - Huge amounts of **hand-crafted features**
  - Reliance on **human curated resources** like dictionaries
  - Most of the above are **language-pair-specific**, have to be repeated
- MT was one of the **first major success stories** of neural methods in NLP:
  - End-to-end systems, “language-agnostic” models, equal/better performance

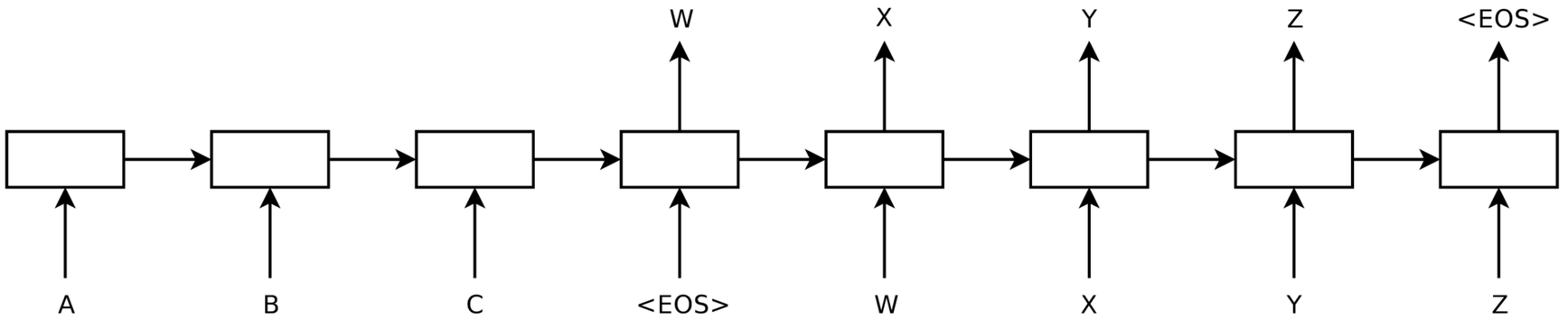
Attention

seq2seq architecture: problem



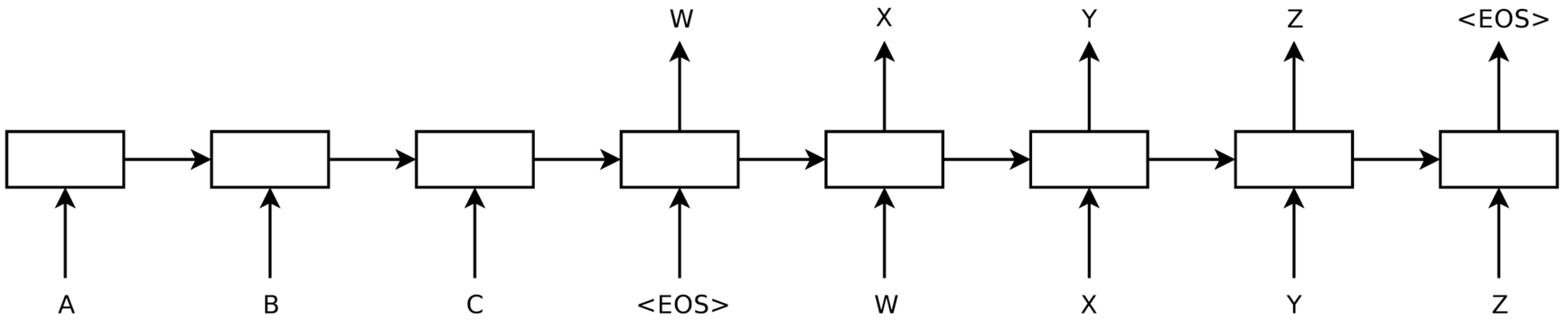
Sutskever et al 2013

seq2seq architecture: problem



Sutskever et al 2013

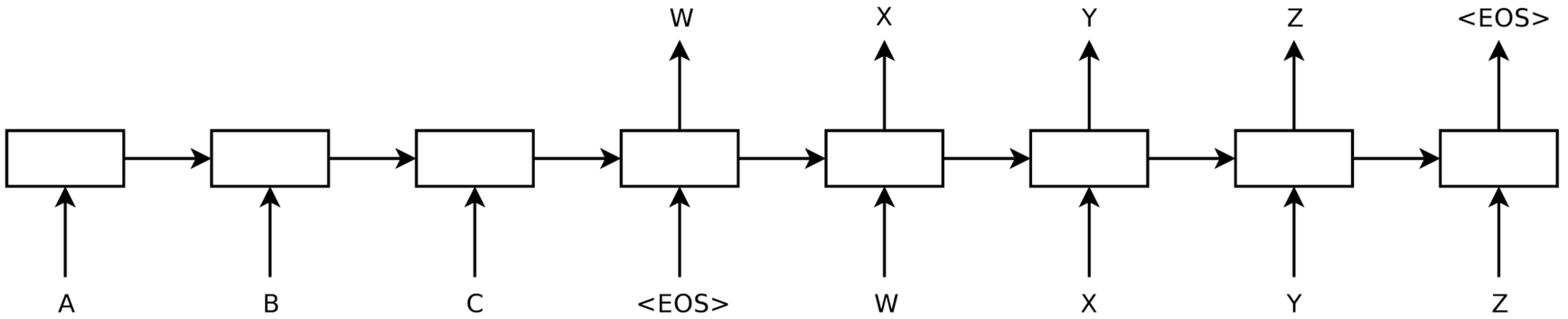
seq2seq architecture: problem



Sutskever et al 2013

seq2seq architecture: problem

Mooney 2014: “You can't cram the meaning of a whole %&!$# sentence into a single $&!#* vector!”

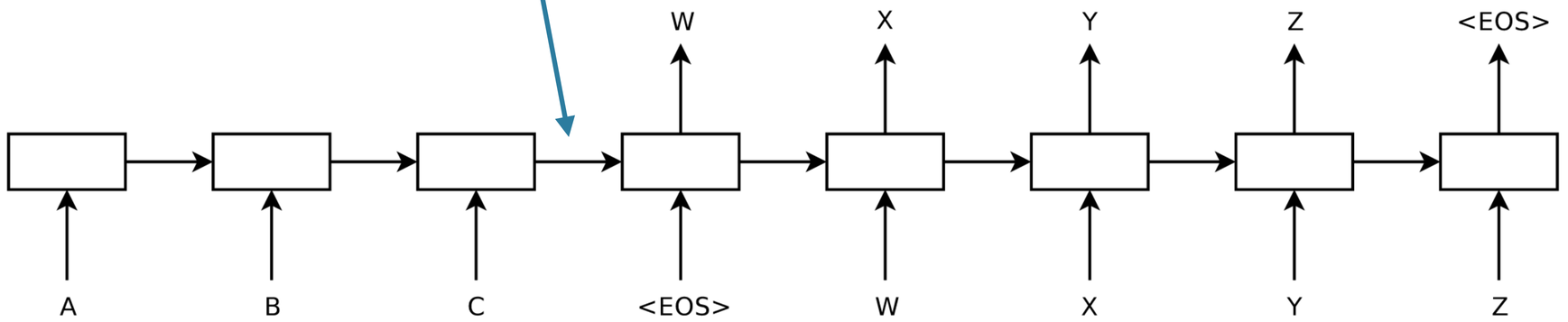


Sutskever et al 2013

seq2seq architecture: problem

Decoder can only see info in this

Mooney 2014: “You can't cram the meaning of a whole %&!$# sentence into a single $&!#* vector!”



Sutskever et al 2013

NEURAL MACHINE TRANSLATION BY JOINTLY LEARNING TO ALIGN AND TRANSLATE

Dzmitry Bahdanau
Jacobs University Bremen, Germany

KyungHyun Cho **Yoshua Bengio***
Université de Montréal

ABSTRACT

Neural machine translation is a recently proposed approach to machine translation. Unlike the traditional statistical machine translation, the neural machine translation aims at building a single neural network that can be jointly tuned to maximize the translation performance. The models proposed recently for neural machine translation often belong to a family of encoder–decoders and encode a source sentence into a fixed-length vector from which a decoder generates a translation. In this paper, we conjecture that the use of a fixed-length vector is a bottleneck in improving the performance of this basic encoder–decoder architecture, and propose to extend this by allowing a model to automatically (soft-)search for parts of a source sentence that are relevant to predicting a target word, without having to form these parts as a hard segment explicitly. With this new approach, we achieve a translation performance comparable to the existing state-of-the-art phrase-based system on the task of English-to-French translation. Furthermore, qualitative analysis reveals that the (soft-)alignments found by the model agree well with our intuition.

source

NEURAL MACHINE TRANSLATION BY JOINTLY LEARNING TO ALIGN AND TRANSLATE

Dzmitry Bahdanau
Jacobs University Bremen, Germany

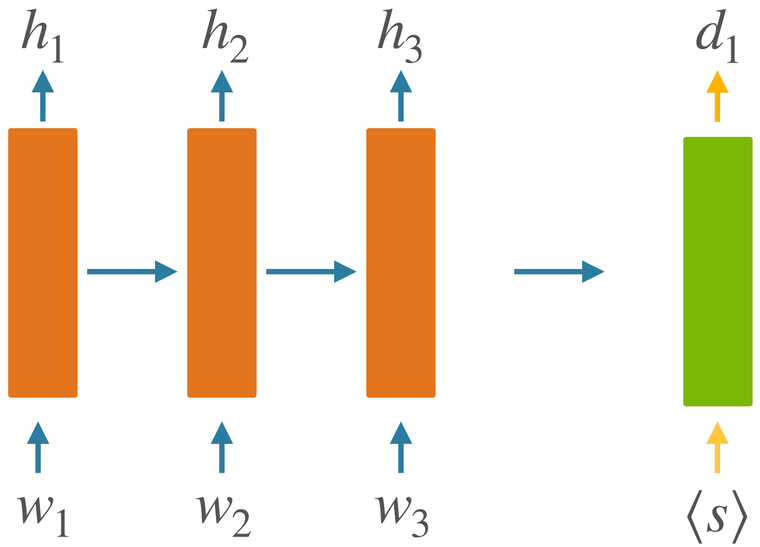
KyungHyun Cho **Yoshua Bengio***
Université de Montréal

ABSTRACT

Neural machine translation is a recently proposed approach to machine translation. Unlike the traditional statistical machine translation, the neural machine translation aims at building a single neural network that can be jointly tuned to maximize the translation performance. The models proposed recently for neural machine translation often belong to a family of encoder–decoders and encode a source sentence into a fixed-length vector from which a decoder generates a translation. In this paper, we conjecture that the use of a fixed-length vector is a bottleneck in improving the performance of this basic encoder–decoder architecture, and propose to extend this by allowing a model to automatically (soft-)search for parts of a source sentence that are relevant to predicting a target word, without having to form these parts as a hard segment explicitly. With this new approach, we achieve a translation performance comparable to the existing state-of-the-art phrase-based system on the task of English-to-French translation. Furthermore, qualitative analysis reveals that the (soft-)alignments found by the model agree well with our intuition.

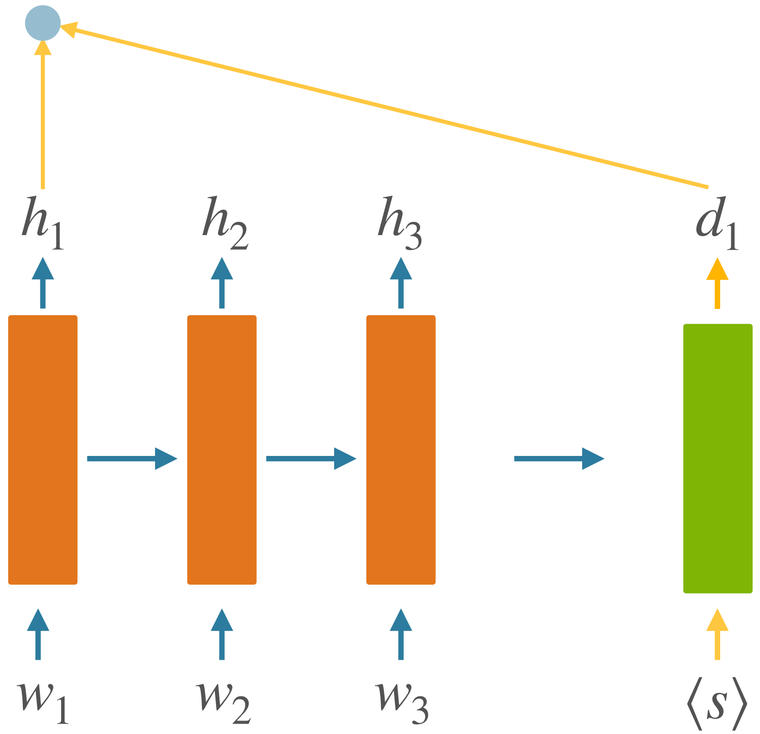
source

Adding Attention



Badhanau et al 2014
Luong et al 2015

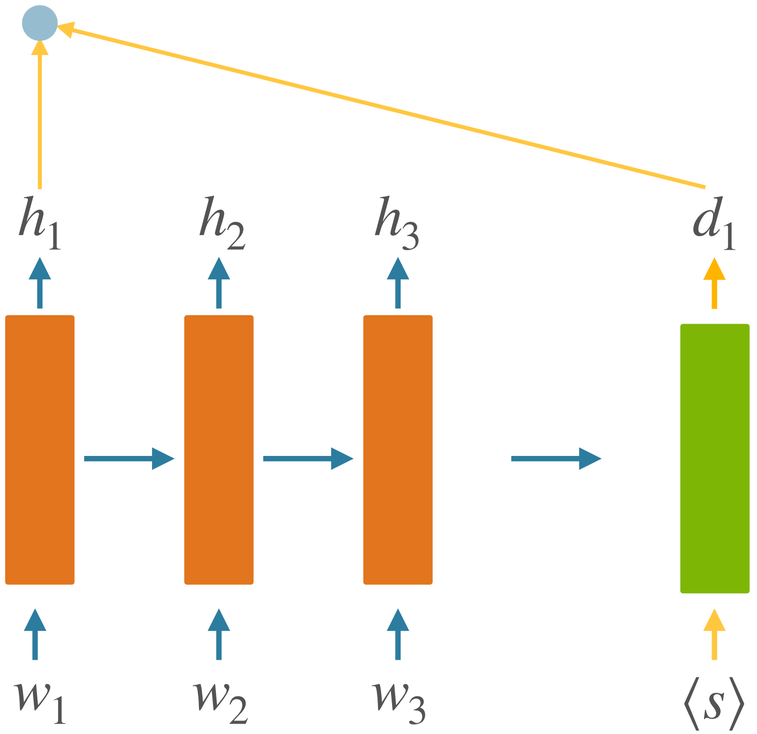
Adding Attention



Badhanau et al 2014
Luong et al 2015

Adding Attention

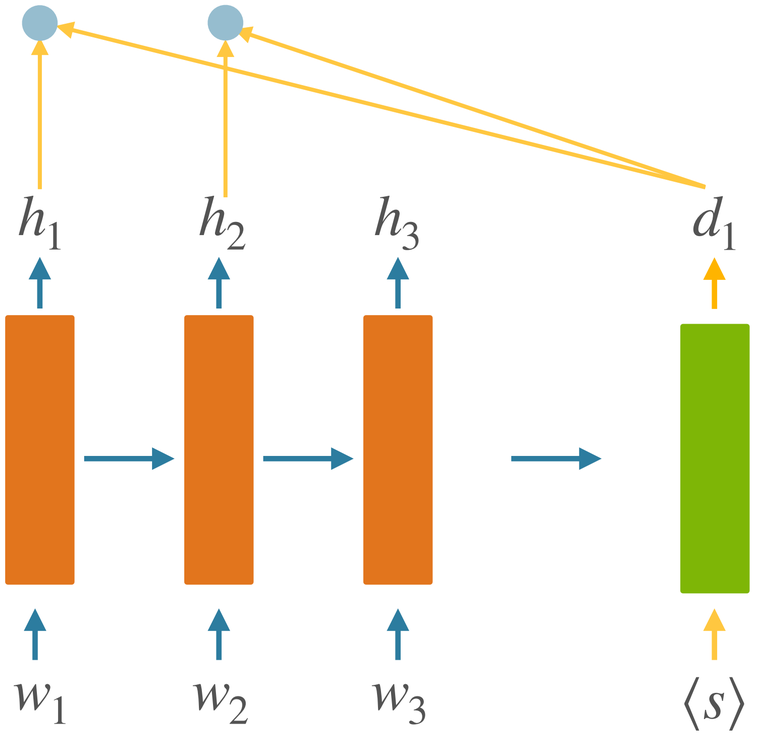
$\alpha_{ij} = a(h_j, d_i)$
(dot product usually)



Badhanau et al 2014
Luong et al 2015

Adding Attention

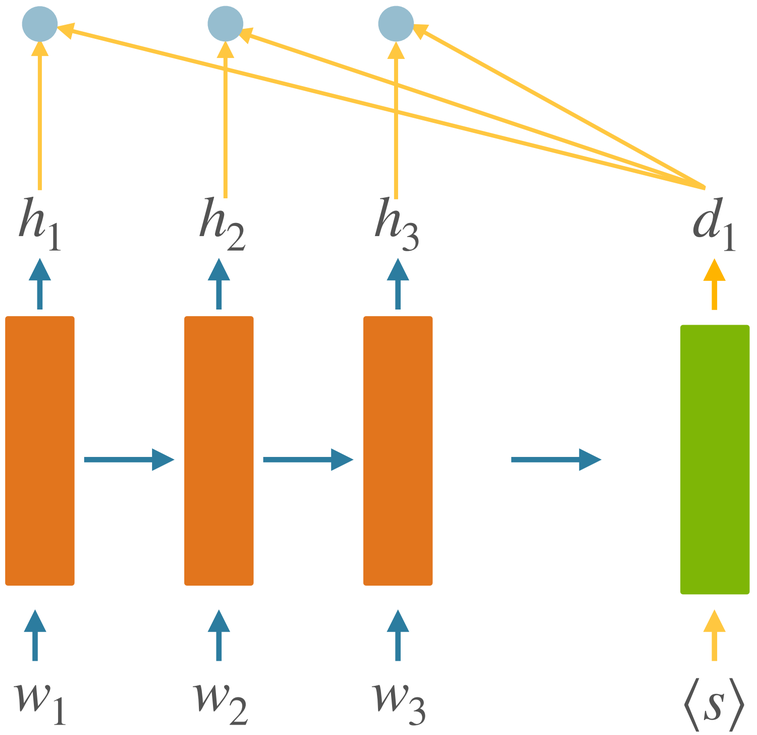
$\alpha_{ij} = a(h_j, d_i)$
(dot product usually)



Badhanau et al 2014
Luong et al 2015

Adding Attention

$\alpha_{ij} = a(h_j, d_i)$
(dot product usually)

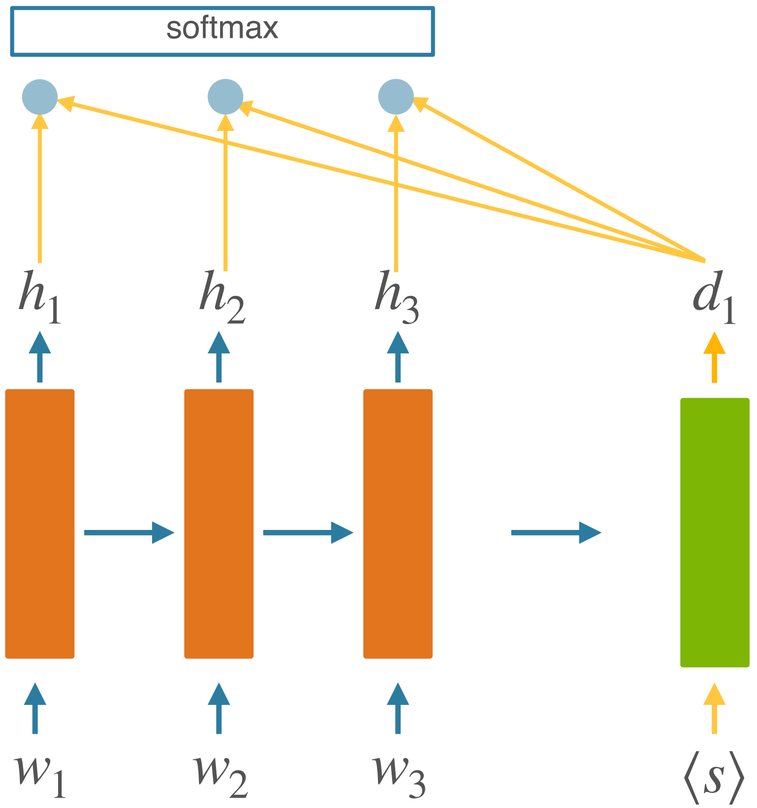


Badhanau et al 2014
Luong et al 2015

Adding Attention

$e_{ij} = \text{softmax}(\alpha)_j$

$\alpha_{ij} = a(h_j, d_i)$
(dot product usually)



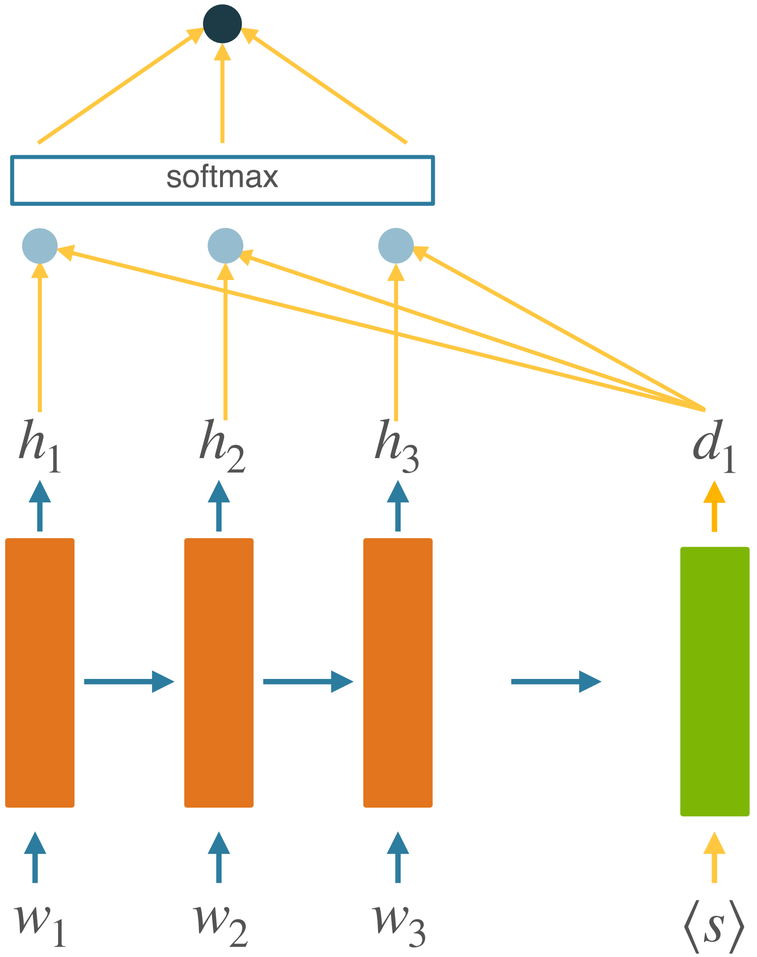
Badhanau et al 2014
Luong et al 2015

Adding Attention

$c_i = \Sigma_j e_{ij} h_j$

$e_{ij} = \text{softmax}(\alpha)_j$

$\alpha_{ij} = a(h_j, d_i)$
(dot product usually)



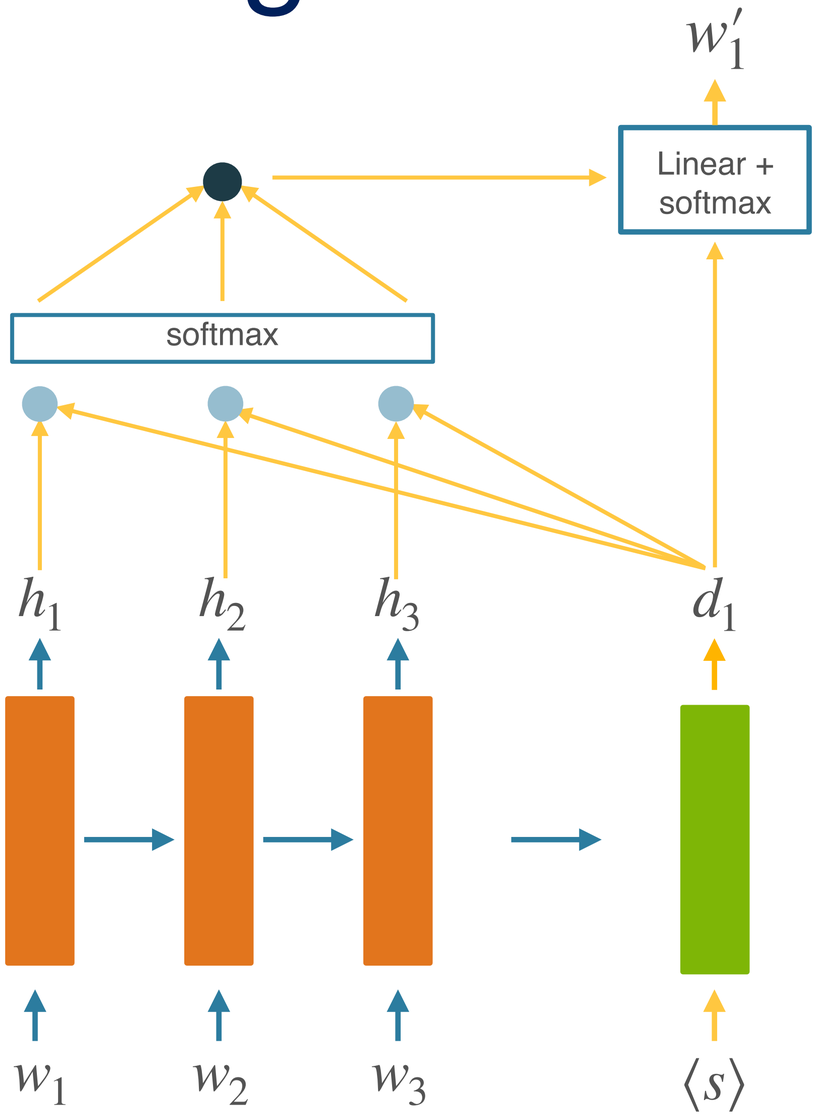
Badhanau et al 2014
Luong et al 2015

Adding Attention

$c_i = \Sigma_j e_{ij} h_j$

$e_{ij} = \text{softmax}(\alpha)_j$

$\alpha_{ij} = a(h_j, d_i)$
(dot product usually)



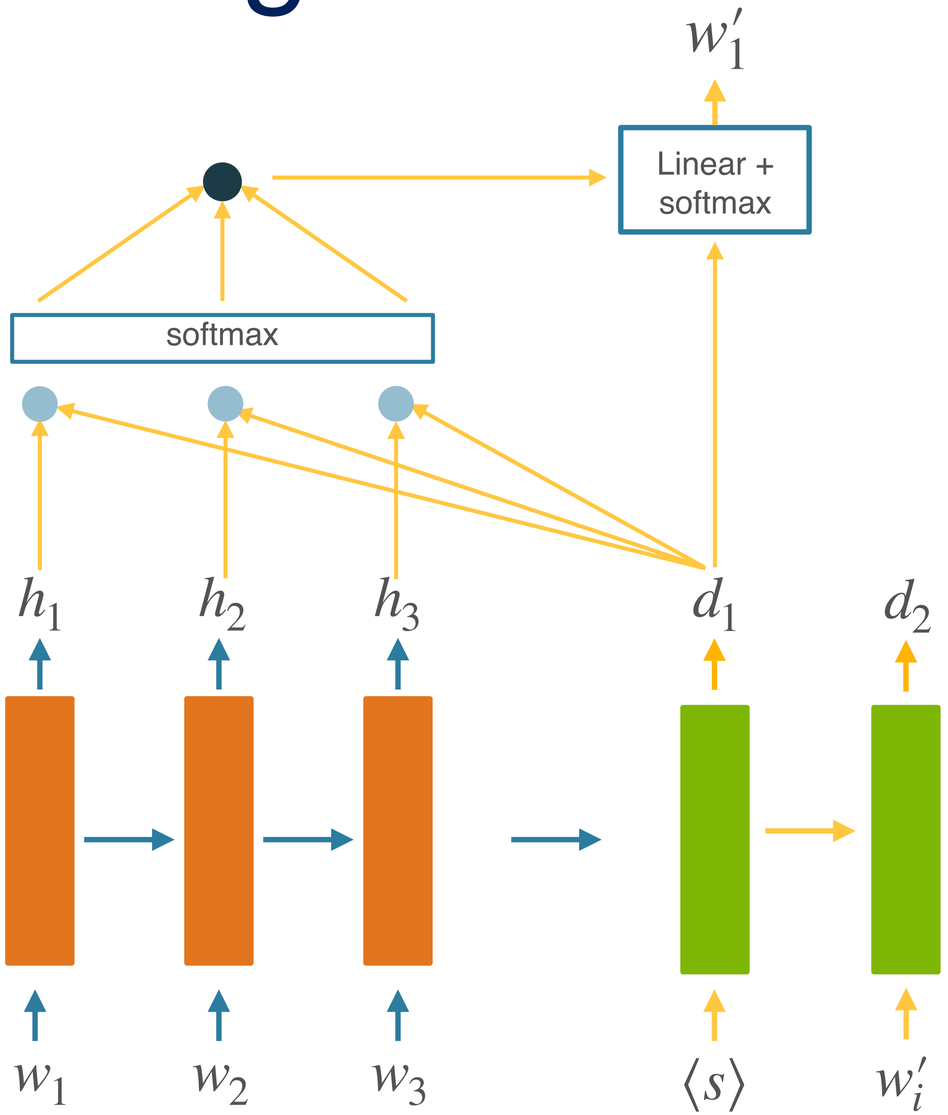
Badhanau et al 2014
Luong et al 2015

Adding Attention

$c_i = \Sigma_j e_{ij} h_j$

$e_{ij} = \text{softmax}(\alpha)_j$

$\alpha_{ij} = a(h_j, d_i)$
(dot product usually)



Badhanau et al 2014
Luong et al 2015

Attention, Generally

Attention, Generally

- A **Query** q pays attention to **Values** $\{v_k\}$ based on **similarity** with **Keys** $\{k_v\}$

Attention, Generally

- A **Query** q pays attention to **Values** $\{v_k\}$ based on **similarity** with **Keys** $\{k_v\}$
- Dot-product attention:

$$\alpha_j = q \cdot k_j$$
$$e_j = e^{\alpha_j} / \Sigma_j e^{\alpha_j}$$
$$c = \Sigma_j e_j v_j$$

Attention, Generally

- A **Query** q pays attention to **Values** $\{v_k\}$ based on **similarity** with **Keys** $\{k_v\}$
- Dot-product attention:

$$\alpha_j = q \cdot k_j$$
$$e_j = e^{\alpha_j} / \Sigma_j e^{\alpha_j}$$
$$c = \Sigma_j e_j v_j$$

- In the previous example: **encoder hidden states** played **both** the keys and the values roles

Why attention?

Why attention?

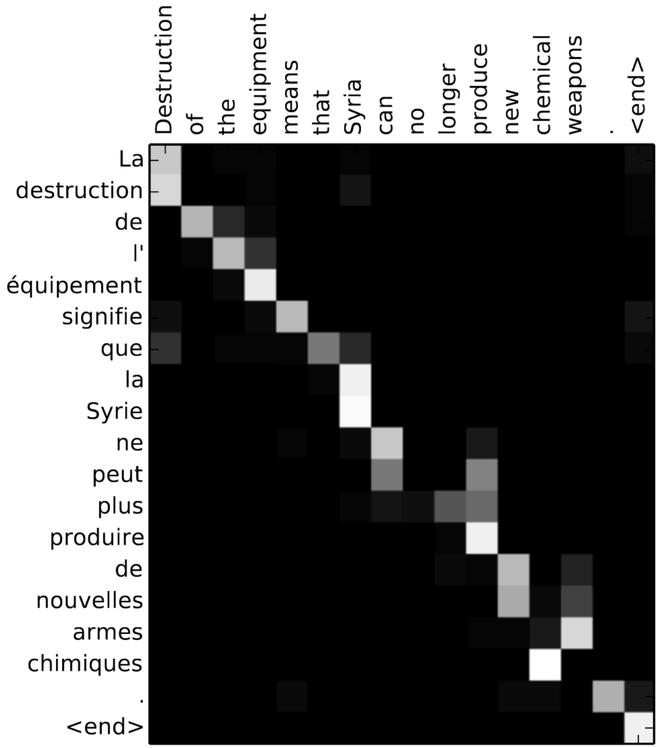
- Incredibly useful (for performance)
  - By **“solving” the bottleneck issue**

Why attention?

- Incredibly useful (for performance)
  - By **“solving” the bottleneck issue**
- Aids **interpretability (maybe)**

Why attention?

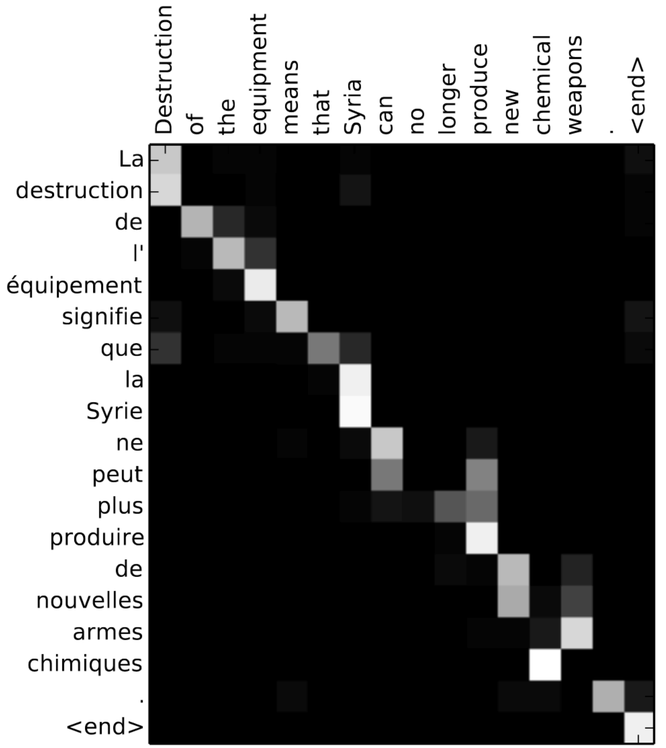
- Incredibly useful (for performance)
  - By **“solving” the bottleneck issue**
- Aids **interpretability (maybe)**



Badhanau et al 2014

Why attention?

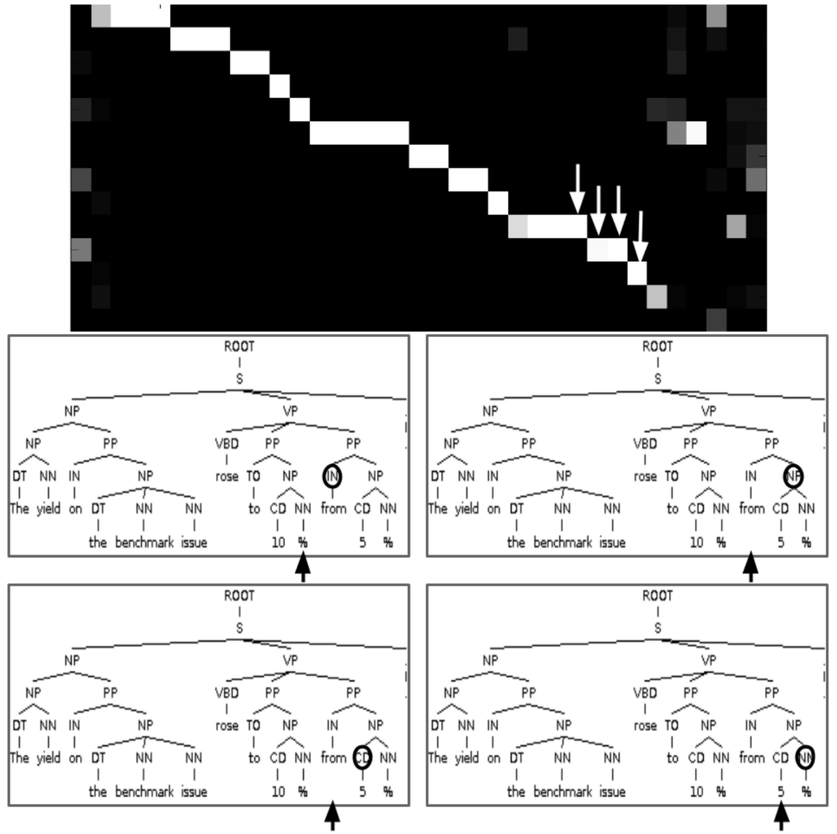
- Incredibly useful (for performance)
  - By **“solving” the bottleneck issue**
- Aids **interpretability (maybe)**
- A **general technique** for combining representations, applications in:
  - NMT, parsing, image/video captioning, more



Badhanau et al 2014

Why attention?

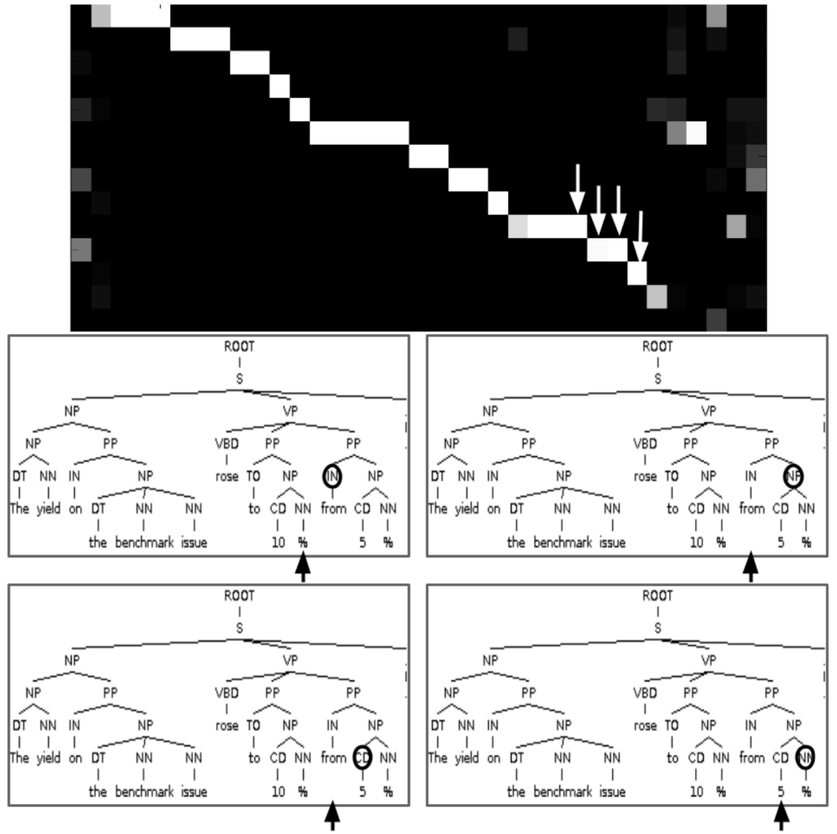
- Incredibly useful (for performance)
  - By **“solving” the bottleneck issue**
- Aids **interpretability (maybe)**
- A **general technique** for combining representations, applications in:
  - NMT, parsing, image/video captioning, more



Vinyals et al 2015

Why attention?

- Incredibly useful (for performance)
  - By **“solving” the bottleneck issue**
- Aids **interpretability (maybe)**
- A **general technique** for combining representations, applications in:
  - NMT, parsing, image/video captioning, more
- Conceptually, let the model **learn to align** representations
  - “Soft” alignment, just like gates = “soft” masks



Vinyals et al 2015