MATHEMATICAL FRONTIERS

The National Academies of SCIENCES ENGINEERING MEDICINE

nas.edu/MathFrontiers

Board on
Mathematical Sciences & Analytics

MATHEMATICAL FRONTIERS

2019 Monthly Webinar Series, 2-3pm ET

February 12: *Machine Learning for Materials Science**

March 12: *Mathematics of Privacy**

April 9: *Mathematics of Gravitational Waves**

May 14: *Algebraic Geometry**

June 11: *Mathematics of Transportation**

July 9: *Cryptography & Cybersecurity**

August 13: *Machine Learning in Medicine**

September 10: *Logic and Foundations**

October 8: *Mathematics of Quantum Physics**

November 12: *Quantum Encryption**

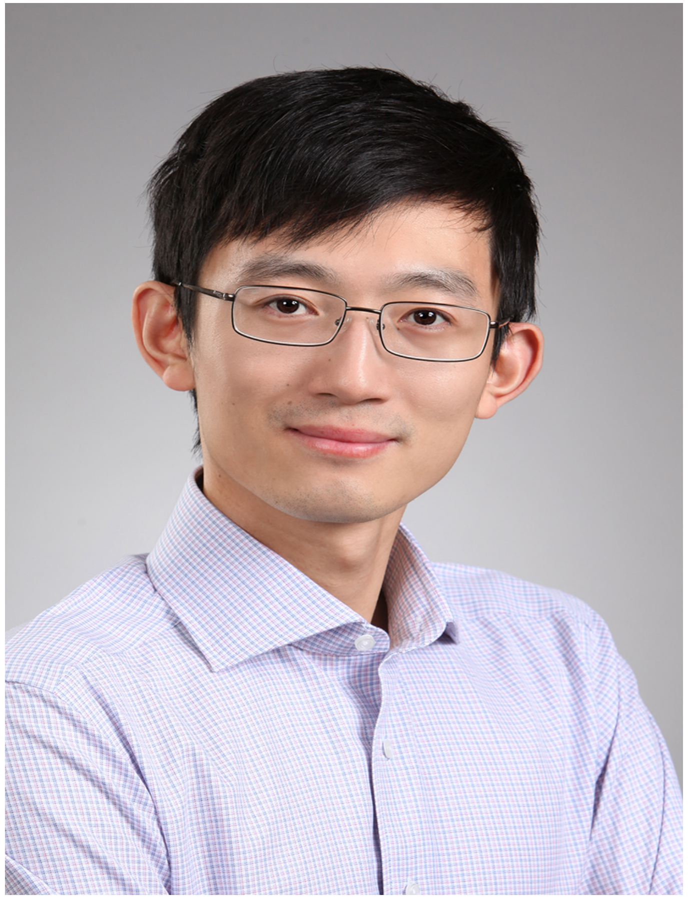
December 10: *Machine Learning for Text*

Made possible by support for BMSA from the
National Science Foundation
Division of Mathematical Sciences
and the
Department of Energy
Advanced Scientific Computing Research

** Webinar posted*

MATHEMATICAL FRONTIERS

Machine Learning for Text



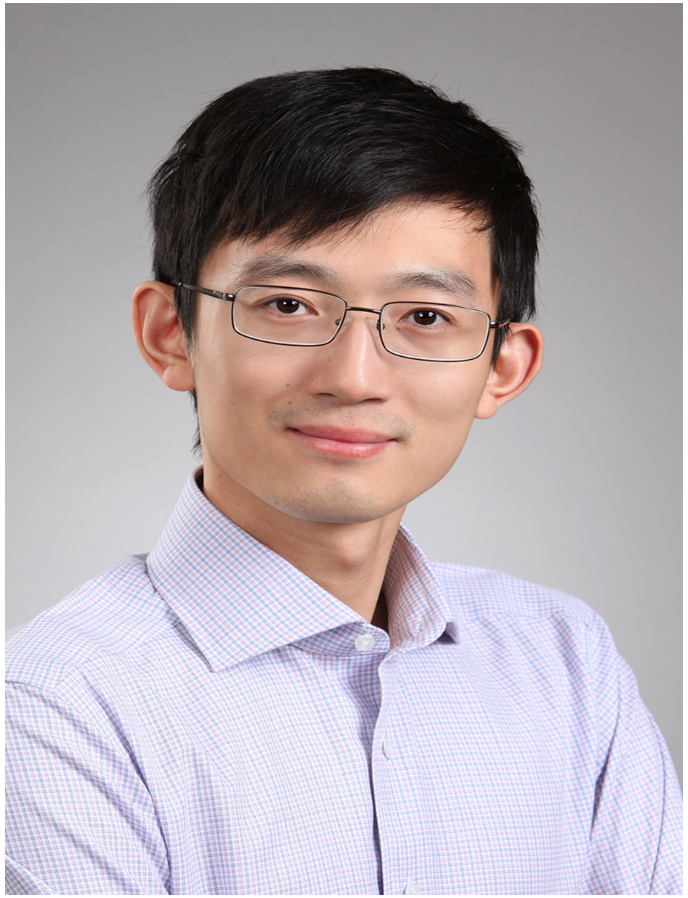
Tengyu Ma,
Stanford University

Marine Carpuat,
University of Maryland

Mark Green,
UCLA (moderator)

MATHEMATICAL FRONTIERS

Machine Learning for Text

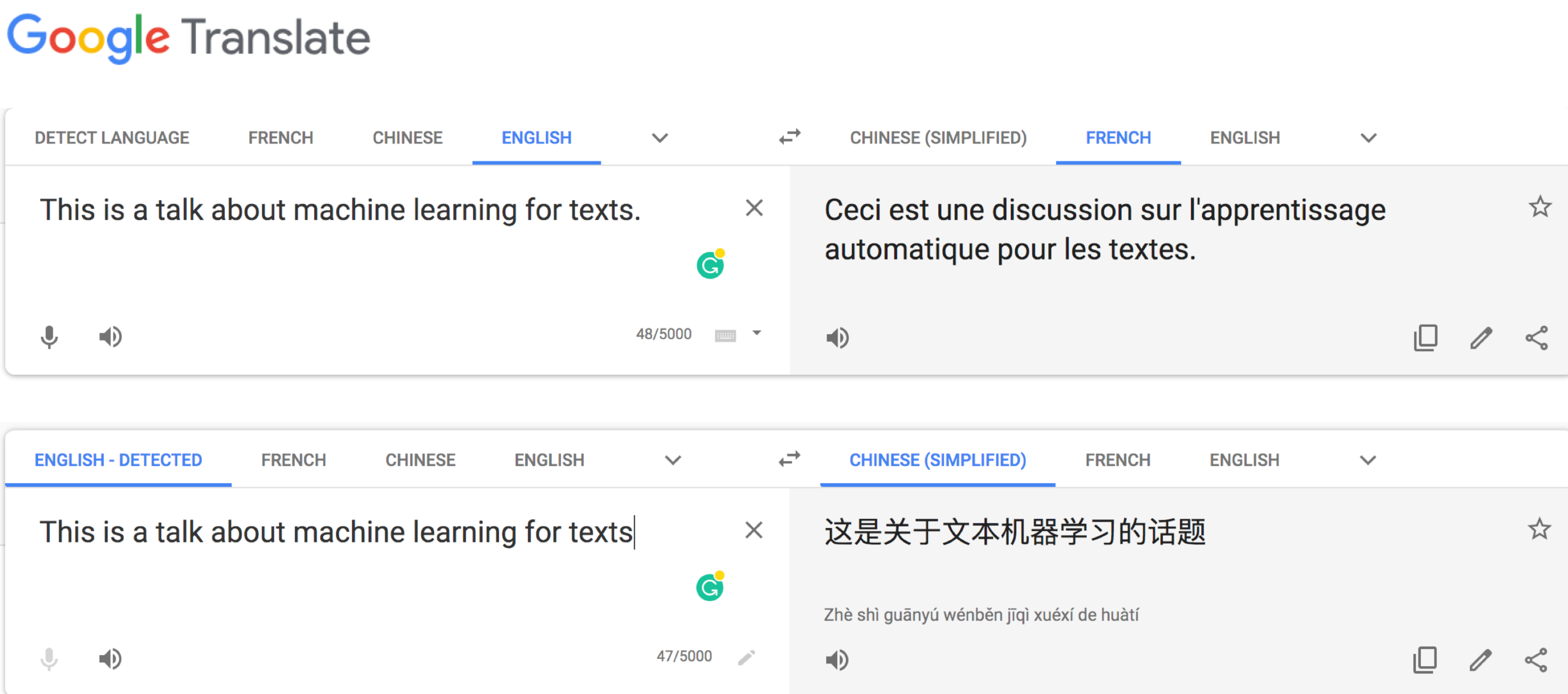


Tengyu Ma,
Stanford University

Assistant Professor of Computer Science and Statistics

Machine Learning for Texts: Understanding Embeddings

Breakthroughs in Natural Language Processing

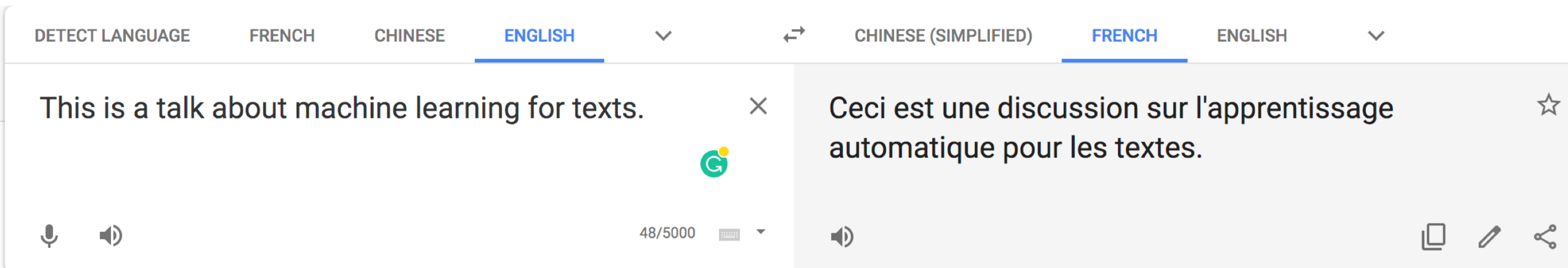


Machine Learning (Supervised Learning)

- Find a function f

Input $\xrightarrow{f}$ output

- Translation



- Sentiment analysis

This seminar series is fantastic.

positive sentiment

Machine Learning (Supervised Learning)

- Find a function f

$$\text{Input} \xrightarrow{\quad f \quad} \text{output}$$

How do we represent texts inputs as numerical values?

Classic "One-hot" and "Bag-of-words" Representation

- Vocabulary = {a, aardvark, aardwolf, …, zymurgy} of size N

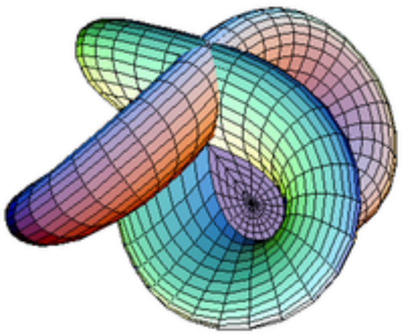
$$
\text{happy} \longrightarrow v_{\text{happy}} = \begin{array}{c} \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ \vdots \\ 1 \\ \vdots \\ \vdots \\ \vdots \\ 0 \end{bmatrix} \end{array} \begin{array}{c} \text{a} \\ \text{aardvark} \\ \text{aardwolf} \\ \\ \\ \text{happy} \\ \\ \\ \\ \text{zymurgy} \end{array} \in \mathbb{R}^N
$$

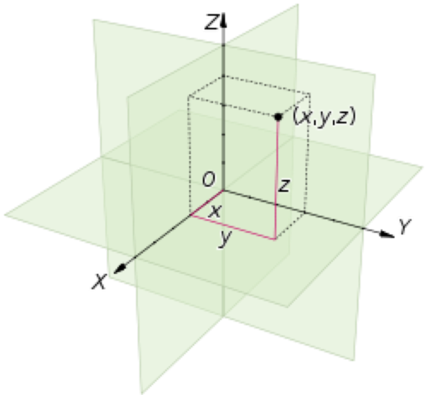
Classic "One-hot" and "Bag-of-words" Representation

- Vocabulary = {a, aardvark, aardwolf, …, zymurgy} of size N

$$\text{A happy aardwolf} \longrightarrow v = \begin{bmatrix}1\\0\\1\\\vdots\\\vdots\\1\\\vdots\\\vdots\\\vdots\\0\end{bmatrix} \begin{array}{l}\text{a}\\\text{aardvark}\\\text{aardwolf}\\ \\ \\\text{happy}\\ \\ \\ \\\text{zymurgy}\end{array} \in \mathbb{R}^N$$

Embeddings (in Machine Learning)





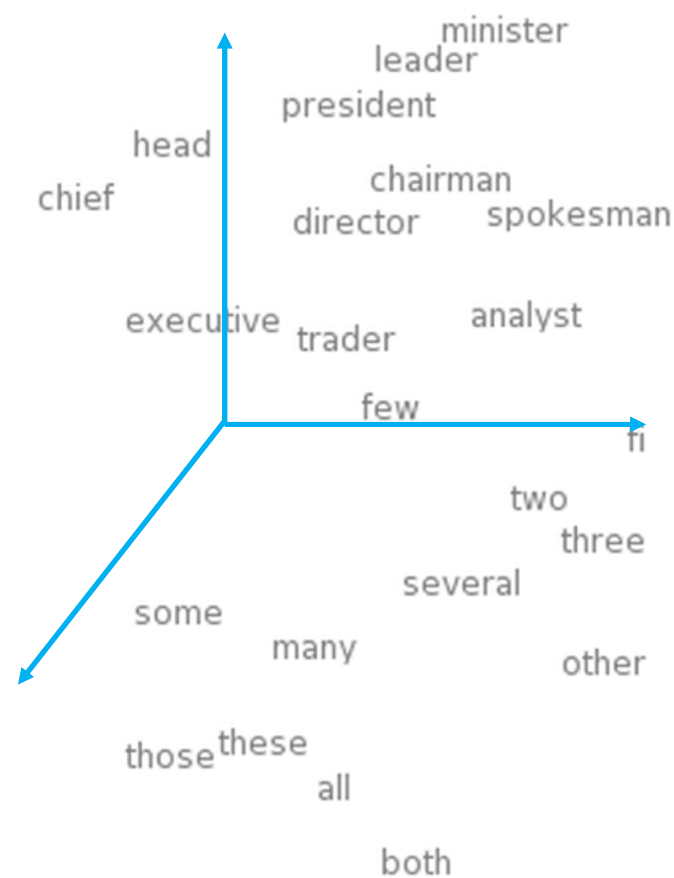
$x \in \mathcal{X}$ $\longrightarrow$ $v_x \in \mathbb{R}^d$

complicated space

Euclidean space with meaningful inner products

Word Embeddings

Vocabulary $\longrightarrow$ $\mathbb{R}^{300}$



Goal:

embedding captures semantics information

(ideally via linear algebraic operations)

- inner products characterize similarity
  - similar words have large inner products
- differences characterize relationship
  - analogous pairs have similar differences

Picture credit: Chris Olah's blog post

Distributional Hypothesis of Meaning ([Harris'54], [Firth'57])

Meaning of a word is determined by words it co-occurs with.

$$\begin{array}{cc} & \text{word } y \\ & \downarrow \\ \text{word } x \rightarrow & \begin{bmatrix} & \cdots & \\ \vdots & \boxed{v_x} & \vdots \\ \vdots & & \vdots \\ & \cdots & \end{bmatrix}\end{array}$$

Co-occurrence matrix
$\Pr(\cdot,\cdot)$

Def: $\Pr(x, y) \triangleq$ prob. of co-occurrences of $\underbrace{x, y \text{ in a window of size 5}}_{\text{“a window of size 5”}}$

- Rows of co-occurrence matrix are reasonable embeddings [Lund-Burgess'96]
- [Church-Hanks'90]

v_x = row of $\mathrm{PMI}(x, y) \triangleq \log \frac{\Pr[x,y]}{\Pr[x]\Pr[y]}$

(PMI = point-wise mutual information)

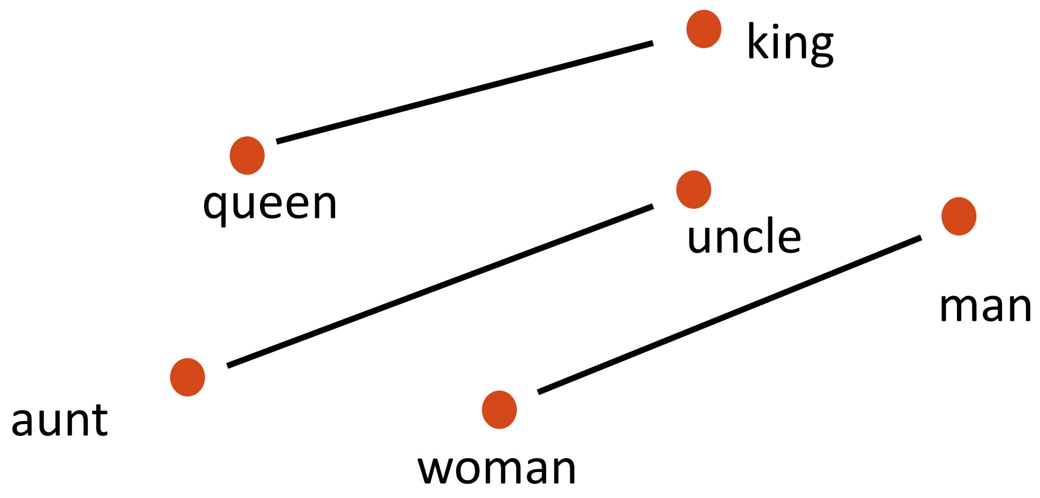
Dimension-Reduced PMI Embeddings [Levy-Goldberg'14]

1. Compute $\text{PMI}(x, y) = \log \frac{\Pr[x,y]}{\Pr[x]\Pr[y]}$
2. Take rank-300 SVD (best rank-300 approximation) of PMI

   ➢ $\Leftrightarrow$ Fit $\text{PMI}(x, y) \approx \langle v_x, v_y \rangle$ where $v_x \in \mathbb{R}^{300}$

➢ "Linear structure" in the found v_x's :

$$v_{woman} - v_{man} \approx v_{queen} - v_{king} \approx v_{uncle} - v_{aunt} \approx \cdots$$



Non-linear Embedding methods

- word2vec [Mikolov et al'13] :

$$\Pr[\, x_{i+6} \mid x_{i+1}, \dots, x_{i+5} \,] \propto \exp\langle v_{x_{i+6}}, \frac{1}{5}\left(v_{x_{i+1}} + \cdots + v_{x_{i+5}}\right)\rangle$$

- GloVe [Pennington et al'14] :

$$\log \Pr[x, y] \approx \langle v_x, v_y \rangle + s_x + s_y + C$$

- [Levy-Goldberg'14] (Previous slide)

$$\mathrm{PMI}(x, y) = \log \frac{\Pr[x,y]}{\Pr[x] \Pr[y]} \approx \langle v_x, v_y \rangle + C$$

Logarithm (or exponential) seems to exclude linear algebra!

Where does the log come from?

[Arora et al.'16, c.f. Levy-Goldberg'14, Pennington et al'14]

- For most of the words χ:

$$\frac{\Pr[\chi \mid king]}{\Pr[\chi \mid queen]} \approx \frac{\Pr[\chi \mid man]}{\Pr[\chi \mid woman]}$$

  - For χ unrelated to gender: LHS, RHS ≈ 1
  - for χ =dress, LHS, RHS $\ll 1$; for χ = John, LHS, RHS $\gg 1$

$$\Rightarrow \sum_{\chi} \left(\log \frac{\Pr[\chi \mid king]}{\Pr[\chi \mid queen]} - \log \frac{\Pr[\chi \mid man]}{\Pr[\chi \mid woman]} \right)^2$$

$$|| \operatorname{PMI}(king, \cdot) - \operatorname{PMI}(queen, \cdot) - \operatorname{PMI}(man, \cdot) + \operatorname{PMI}(woman, \cdot)||_2^2 \approx 0$$

- Rows of PMI matrix has "linear structure"

Why Does Dimension Reduction Help?

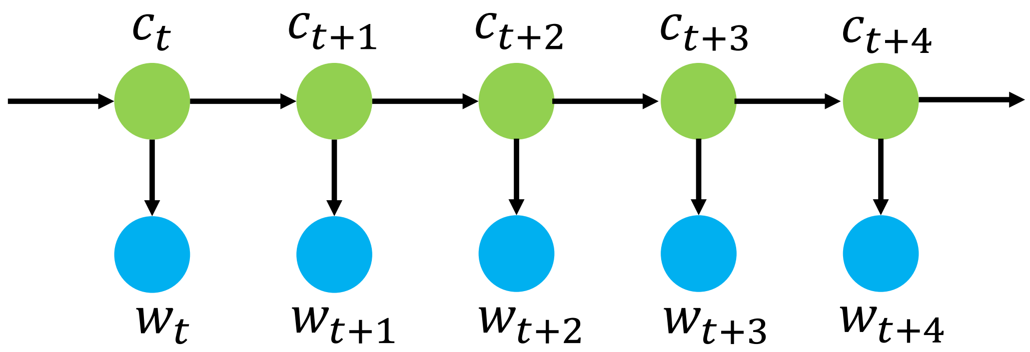
Empirically can find vectors v_x's such that

$$\mathrm{PMI}(x, y) \approx \langle v_x, v_y \rangle$$

1. PMI is not necessarily PSD

2. Relative approximation error is high (17%); Low-dimensional v_x's have better linear structure than rows of PMI

RAND-WALK: A Generative Model for Language [Arora et al'16]



- Hidden Markov Model:
  - discourse vector $c_t \in \mathbb{R}^d$ governs the discourse/theme/context of time t
  - words w_t (observable); embedding $v_{w_t} \in \mathbb{R}^d$ (parameters to learn)
  - log-linear observation model

$$\Pr[w_t \mid c_t] \propto \exp\langle v_{w_t}, c_t \rangle$$

Closely related to [Mnih-Hinton'07]

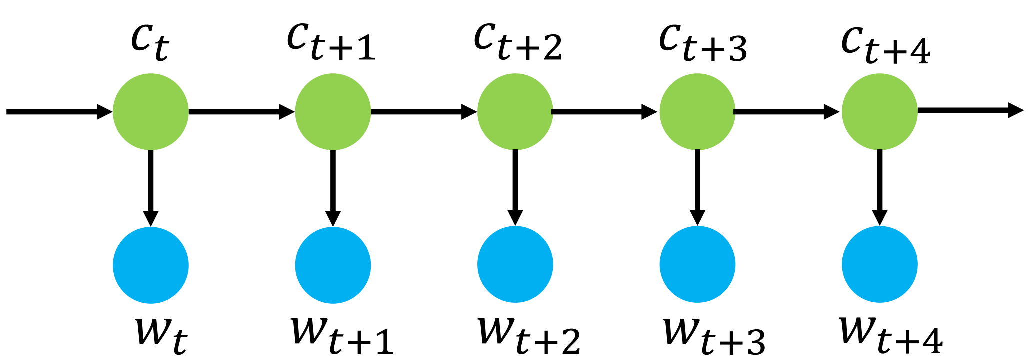
Why Does Dimension Reduction Help?

Empirically can find vectors v_x's such that

$$\mathrm{PMI}(x, y) \approx \langle v_x, v_y \rangle$$

1. PMI is not necessarily PSD and low-rank
   - Under rand-walk model, PMI is approximately PSD and low-rank
2. Relative approximation error is high (17%); Low-dimensional v_x's have better linear structure than rows of PMI
   - Dimension-reduction reduces the noises

RAND-WALK Model



Summary and Looking Ahead

- Theoretical explanations of embeddings methods
  - Popular embeddings methods, such as PMI+SVD, word2vec, Glove can be viewed as algorithms for learning a generative model of language
- Follow-up works: embeddings for sentences, polysemous words, rare words [Arora et al.'17,18a&b …]
- Open directions:
  - Understanding the state-of-the-art contextualized embeddings (Elmo, Bert, etc..)
  - Optimizations of the embeddings
  - Understanding other algorithms for other tasks in NLP (machine translation, etc.)
  - A theory of representation learning

Main References

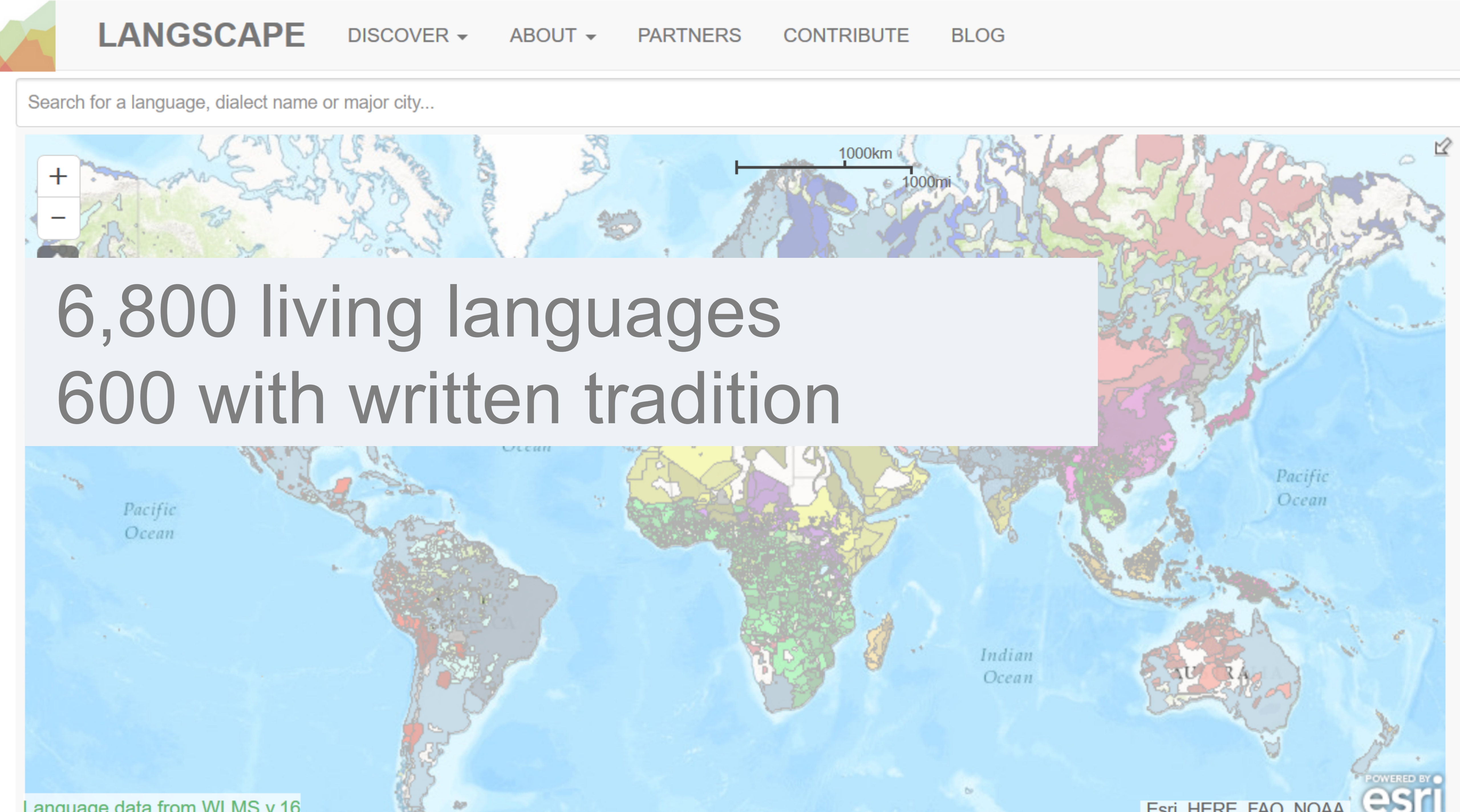
- RAND-WALK: A Latent Variable Model Approach to Word Embeddings. Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. Transactions of the Association for Computational Linguistics (TACL), 2016
- A Simple but Tough-to-Beat Baseline for Sentence Embeddings. Sanjeev Arora, Yingyu Liang, Tengyu Ma. International Conference on Learning Representations (ICLR) 2017
- Linear Algebraic Structure of Word Senses, with Applications to Polysemy. Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. TACL, 2018
- A La Carte Embedding: Cheap but Effective Induction of Semantic Feature Vectors. Mikhail Khodak, Nikunj Saunshi, Yingyu Liang, Tengyu Ma, Brandon Stewart, Sanjeev Arora. ACL, 2018
- Neural Word Embedding as Implicit Matrix Factorization. Omer Levy and Yoav Goldberg. Neurips 2014.
- GloVe: Global Vectors for Word Representation. Jeffrey Pennington, Richard Socher, and Christopher D. Manning. EMNLP, 2014.
- Distributed Representations of Words and Phrases and their Compositionality. Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, Jeffrey Dean. NIPS 2013.

MATHEMATICAL FRONTIERS
Machine Learning for Text

Marine Carpuat,
University of Maryland

Assistant Professor in Computer Science
marine@cs.umd.edu

Toward Breaking Language Barriers with Neural Machine Translation



6,800 living languages
600 with written tradition

UMD Machine Translation
@WMT 2018

这座中国首都拥有速度高得惊人的互联网，有人脸识别软件等尖端技术，在人工智能方面投入了巨资并且拥有无可匹敌的国际化能量，它对富于探索精神的外国人而言是最激动人心的城市之一。

The Chinese capital, with its surprisingly high-speed Internet, sophisticated technology such as face-recognition software, has invested heavily in artificial intelligence and has unrivaled international energy, and is one of the most exciting cities for exploration-minded foreigners.

[Xu & Carpuat WMT 2018]

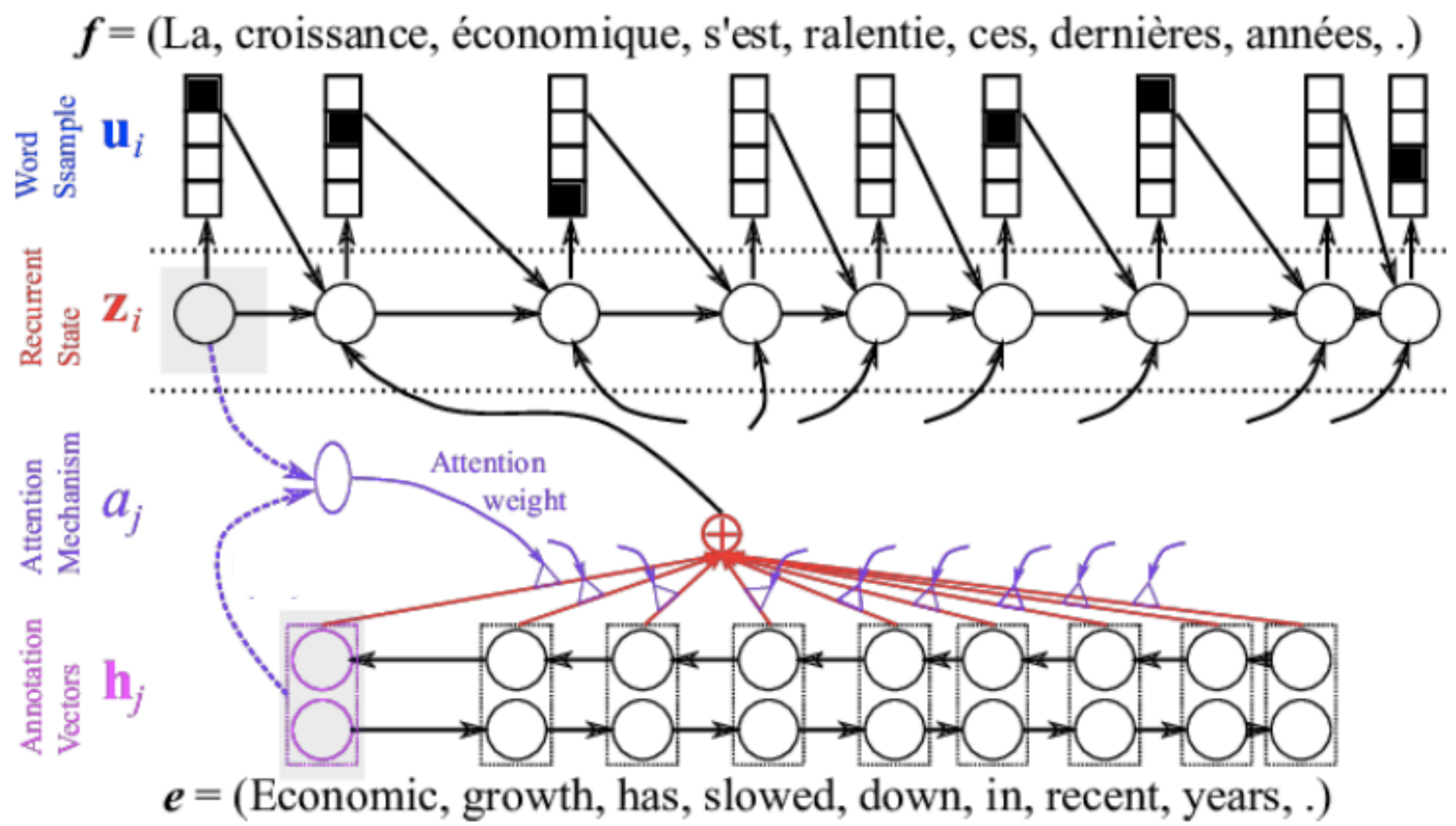
Translation as Machine Learning

An English sentence e is translated into the French sentence

$$f^* = \text{argmax}_f \, p(f \,|\, e; \theta)$$

$$\theta^* = \text{argmax}_\theta \sum_i \log p(f_i \,|\, e_i; \theta)$$

Translation as Deep Learning



$$p(f \mid e;\ \theta) = \prod_{t=1}^{|f|} p(f_t \mid f_{<t}, e;\ \theta)$$

Image: Kyunghung Cho

Translation as Deep Learning: Challenges

requires millions of translation examples

not available for many languages!

raises fundamental machine learning challenges

intractably large output space, infinitely many correct outputs…

makes errors that have real world impact

yet models are opaque, and developed independency from use cases

Toward better translation with limited training data

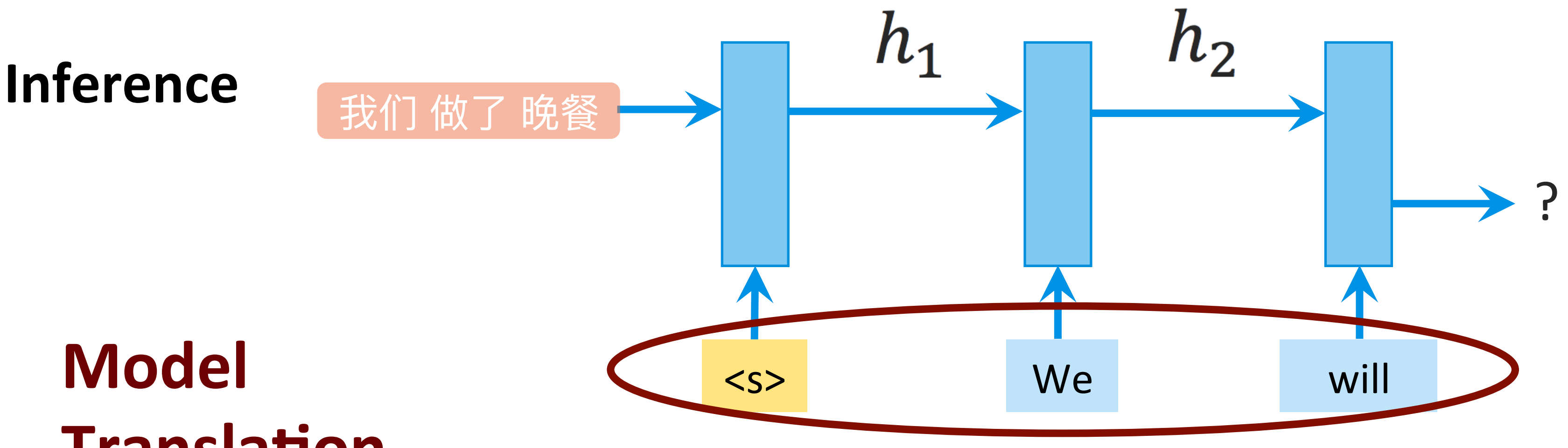
Some approaches:

Learn from related languages

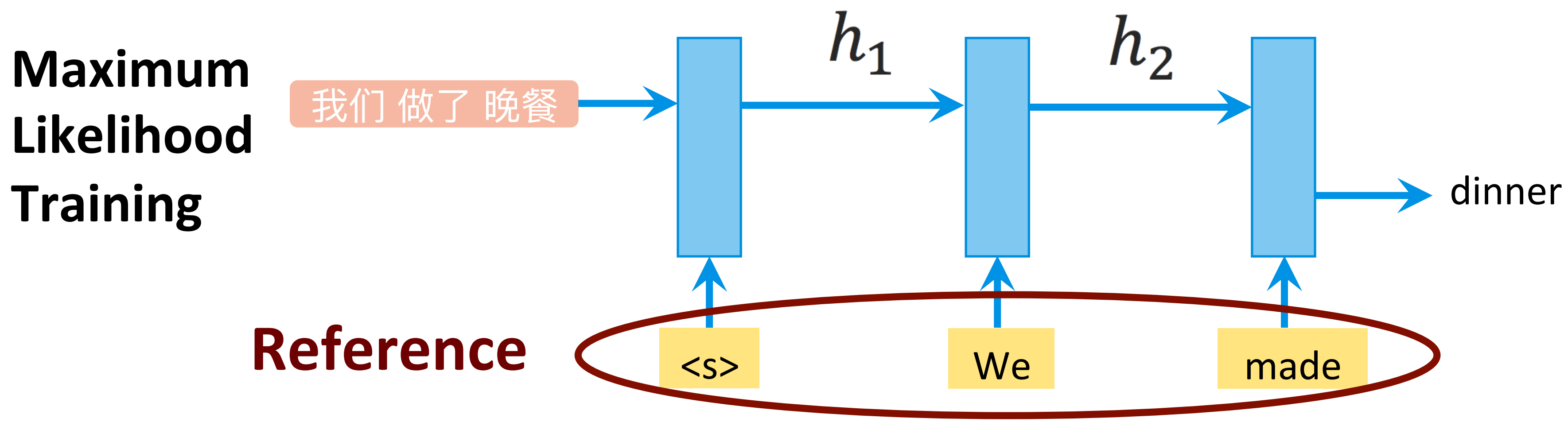
Learn from monolingual text

Improve the training objective

Training Problem: Exposure Bias, a Gap Between Training and Inference



$$P(f|e) = \prod_{t=1}^{T} p(f_t|f_{<t}, e)$$



Loss =

$$\sum_{t=1}^{T} \log p(f_t|f_{<t}, e)$$

How to Address Exposure Bias?

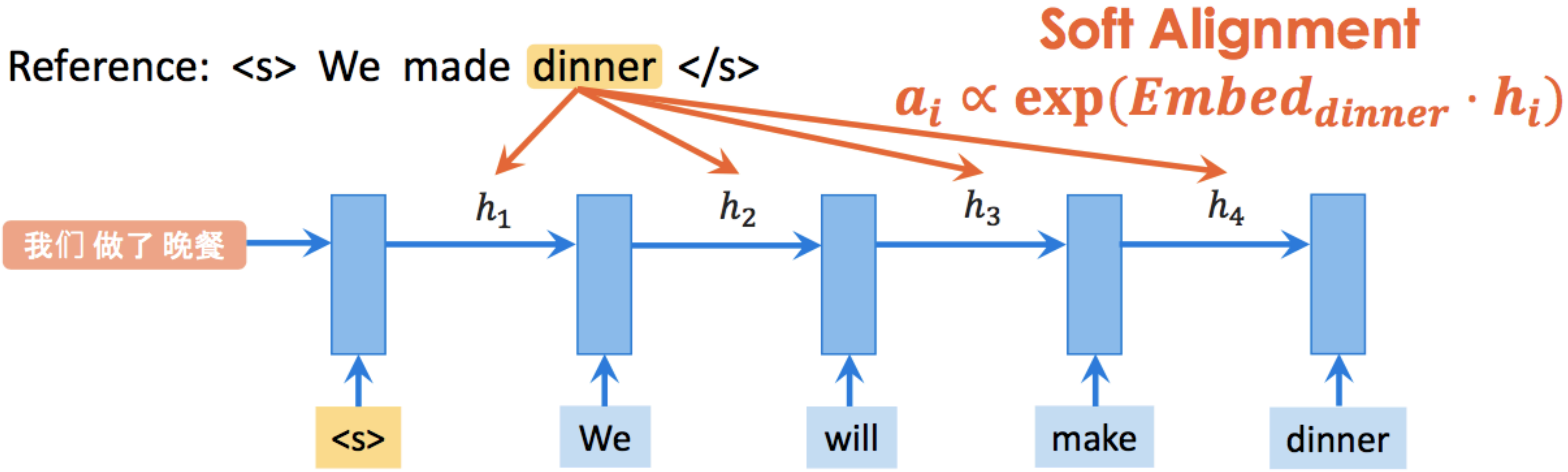
Expose models to their own predictions during training

But how to compute the loss when the partial translation diverges from the reference?

Our method:

1. **Generate translation prefixes** via differentiable sampling
2. Learn to **align** the reference words with sampled prefixes

Our Solution: Align Reference with Partial Translations



a_1 logp(“dinner” | “<s>”, source) + a_2 logp(“dinner” | “<s> We”, source) +
a_3 logp(“dinner” | “<s> We will”, source) + a_4 **logp(“dinner” | “<s> We will make”, source)**

[Xu & Carpuat NAACL 2019]

Toward better translation with limited training data

Some approaches:

Learn from related languages

Learn from monolingual text

Improve the training objective

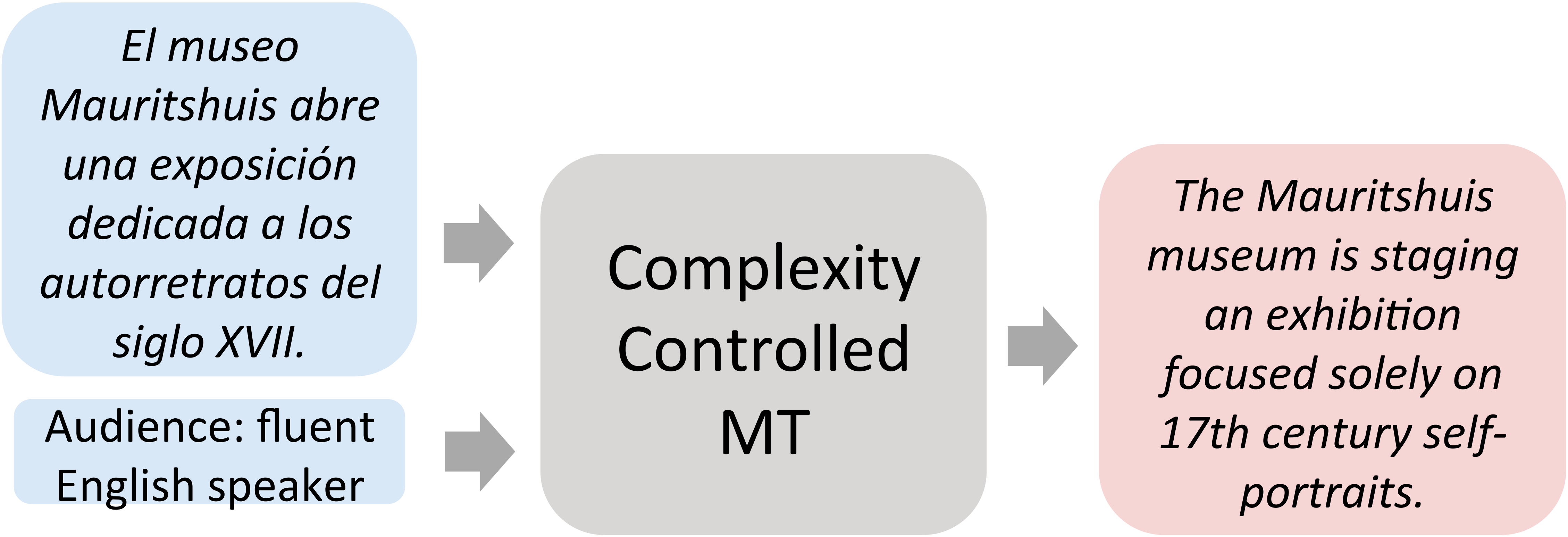
Toward more user-centered machine translation

Can machine translation help human translators and interpreters be more productive?

What errors matter most for different use cases?

Can we tailor machine translation output to different audiences?

Controlling MT Complexity for Different Audiences



Controlling MT Complexity for Different Audiences

El museo Mauritshuis abre una exposición dedicada a los autorretratos del siglo XVII.

Audience: 2nd language learner

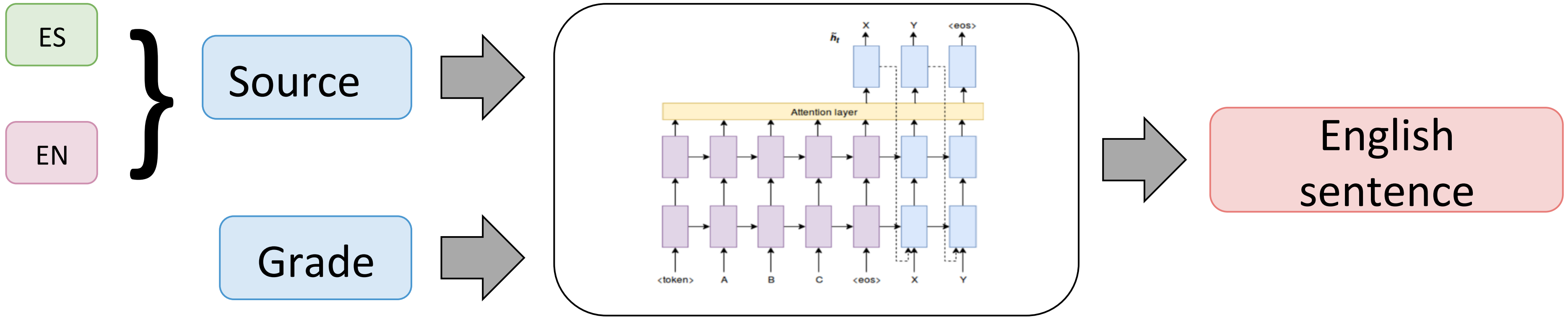
→

Complexity Controlled MT

→

The Mauritshuis museum is going to show self-portraits.

Adapting translation output to different audiences via multi-task learning



$$\text{Multi-task loss} = \underbrace{\sum_{(s_i, g_s, e_o} \log P(e_o | s_i, g_e; \theta)}_{L_{CMT}} + \underbrace{\sum_{(e_i, g_e, e_o} \log P(e_o | e_i, g_e; \theta)}_{L_{Simplify}} + \underbrace{\sum_{(s_i, e_o} \log P(e_o | s_i; \theta)}_{L_{MT}}$$

Spanish sentences translated into simpler English

Complex English sentences paired with simpler English

Spanish-English translation examples

Deep neural networks provide a powerful framework to model translation

How can we improve quality in low-resource settings?

How can we make machine translation more user-centric when it works well?

References

Sweta Agrawal and Marine Carpuat.
“Controlling Text Complexity in Neural Machine Translation”. EMNLP 2019.

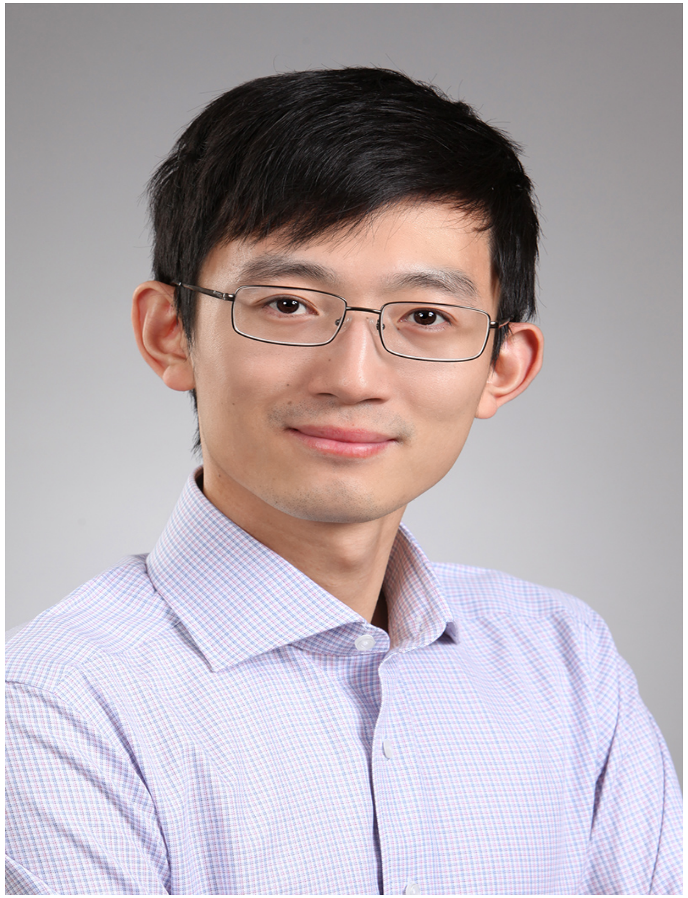
Weijia Xu, Xing Niu and Marine Carpuat.
“Differentiable Sampling with Flexible Reference Word Order for Neural Machine Translation”. NAACL 2019.

Weijia Xu and Marine Carpuat.
“The University of Maryland’s Chinese-English Neural Machine Translation Systems at WMT18”. WMT 2018.

Dzmitry Bahdanau, Kyunghyun Cho, Yoshua Bengio.
“Neural Machine Translation by Jointly Learning to Align and Translate”. ICLR 2015.

MATHEMATICAL FRONTIERS

Machine Learning for Text



Tengyu Ma,
Stanford University

Marine Carpuat,
University of Maryland

Mark Green,
UCLA (moderator)

MATHEMATICAL FRONTIERS

2019 Monthly Webinar Series, 2-3pm ET

February 12: *Machine Learning for Materials Science**

March 12: *Mathematics of Privacy**

April 9: *Mathematics of Gravitational Waves**

May 14: *Algebraic Geometry**

June 11: *Mathematics of Transportation**

July 9: *Cryptography & Cybersecurity**

August 13: *Machine Learning in Medicine**

September 10: *Logic and Foundations**

October 8: *Mathematics of Quantum Physics**

November 12: *Quantum Encryption**

December 10: *Machine Learning for Text*

Made possible by support for BMSA from the
National Science Foundation
Division of Mathematical Sciences
and the
Department of Energy
Advanced Scientific Computing Research

* *Webinar posted*