

291K

Deep Learning for Machine Translation Introduction

Lei Li

UCSB

9/27/2021

Purpose of the Course

- Learn techniques for sequence modeling, structure prediction, and translation with deep learning
- Get familiar with challenges and latest research in machine translation, able to evaluate progress in this area
- Practice engineering skills for building real NLP/MT systems
 - Good for future industry job in ML/NLP
- Get ready to apply to your own current/future research
 - many problems can be formulated as sequence/data to sequence generation

Announcement

- Course website:
 - <https://www.cs.ucsb.edu/~lilei/course/dl4mt21fa/>
- Policy (please read carefully):
 - <https://www.cs.ucsb.edu/~lilei/course/dl4mt21fa/policy.html>
- Reading material
 - Please read before class
- Discussion:
 - <https://piazza.com/class/ksousnwx3cl1ux> (5 for participation in active discussion and sharing)
- Homework
 - HW1-3 separate assignment (10 each)
 - HW4 - In-class presentation: Language in 10 mins (10)
 - HW5 - MT Blog (15)
 - Turn-in your homework at <https://www.gradescope.com/courses/319418>
- Project: proposal, midterm report, final report (40)

Stay Healthy and Safe

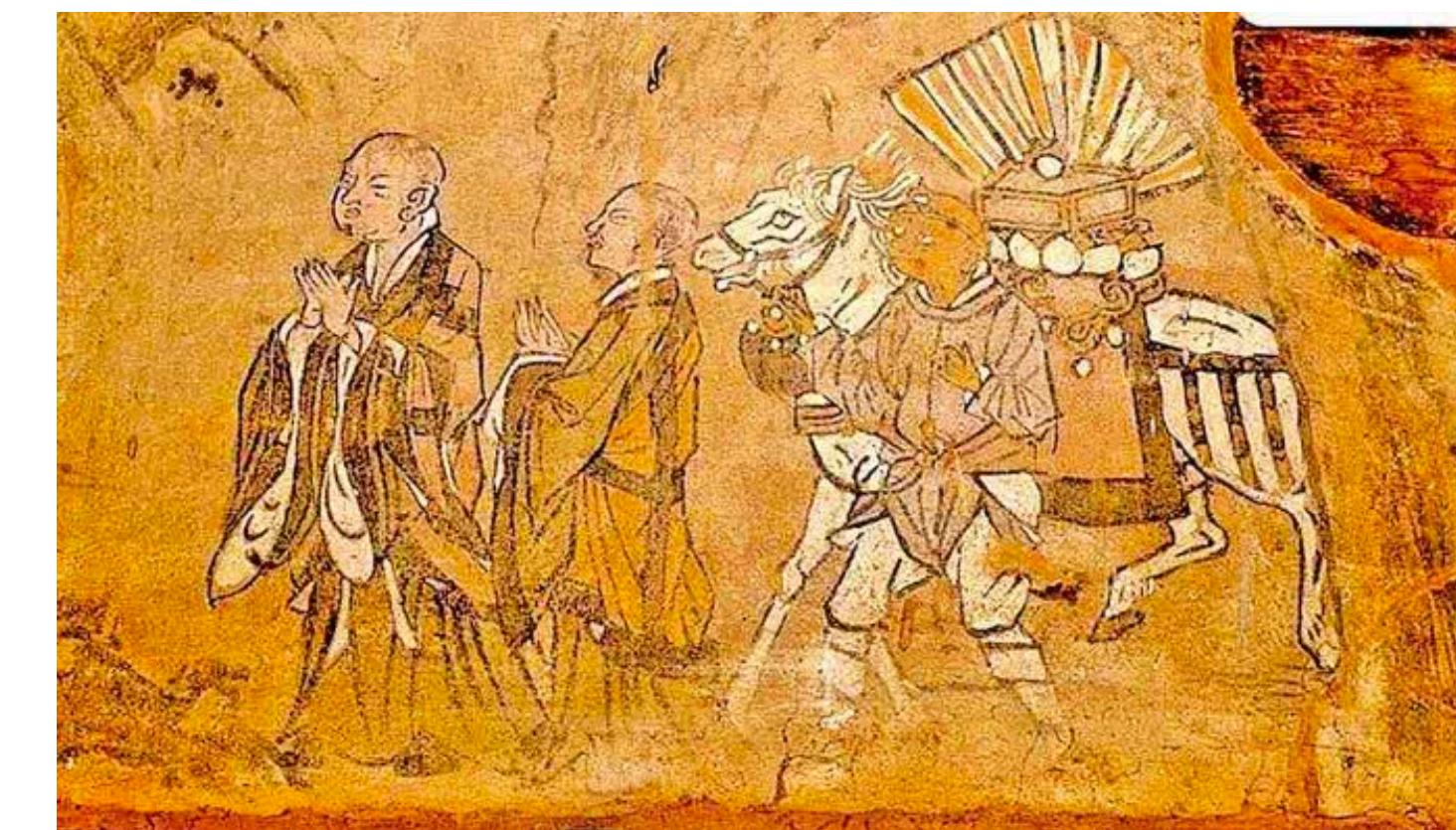
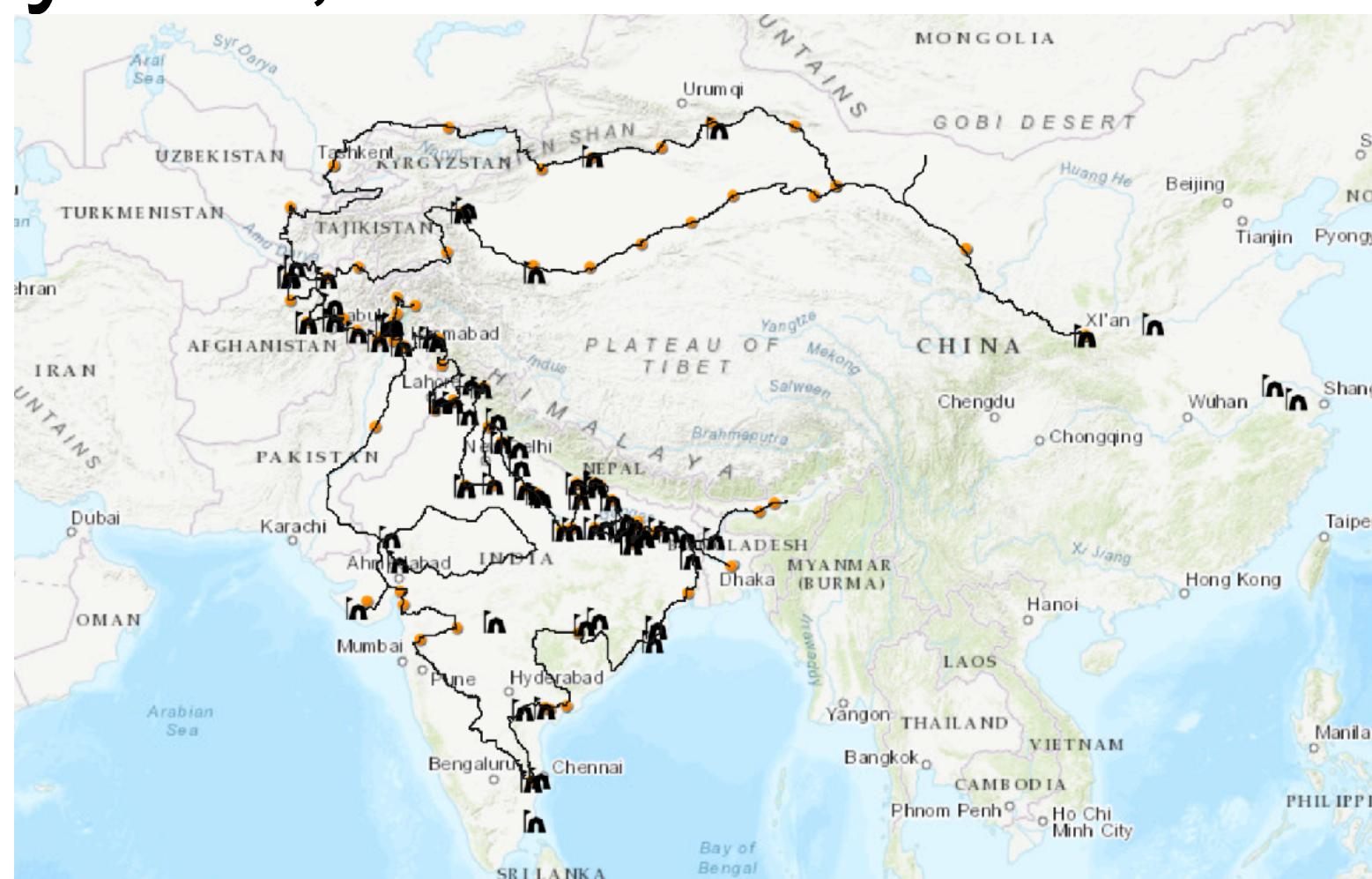
- Adhere to UCSB campus requirement, <https://www.ucsb.edu/COVID-19-information/return-to-campus-requirements>
- Ok to remote attend lectures if not feeling well.

Zoom

- Meeting ID: 821 0655 0534
- Passcode: (shared on Piazza)

Translation originally for Religion need

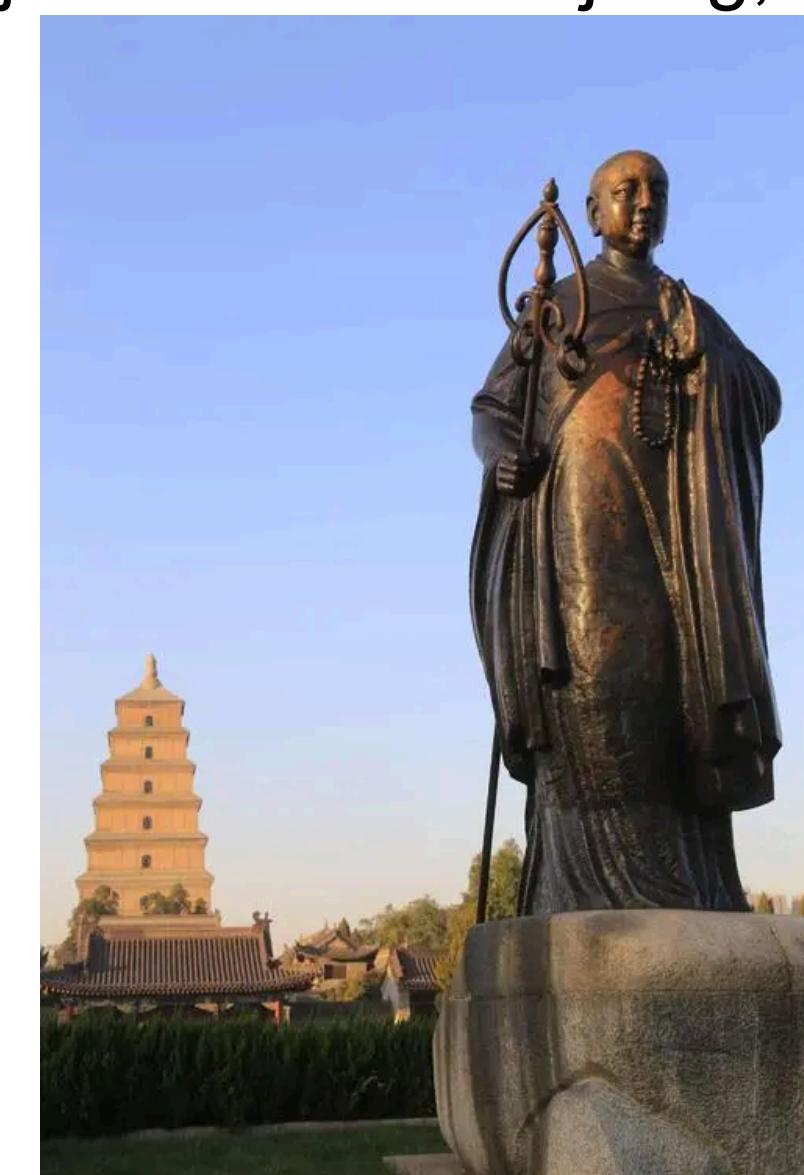
- Septuagint, translated from Hebrew Bible to Greek, mid 3rd century BCE
- Translating Buddhist texts written in Sanskrit to Chinese
 - Kumārajīva (कुमारजीव), 344-413 CE, translated 35-74 books
 - Xuanzang 602-664 CE, travel from China to India in 17 years, translated 75 books from Sanskrit to Chinese



Xuanzang travelling, Dunhuang mural, China



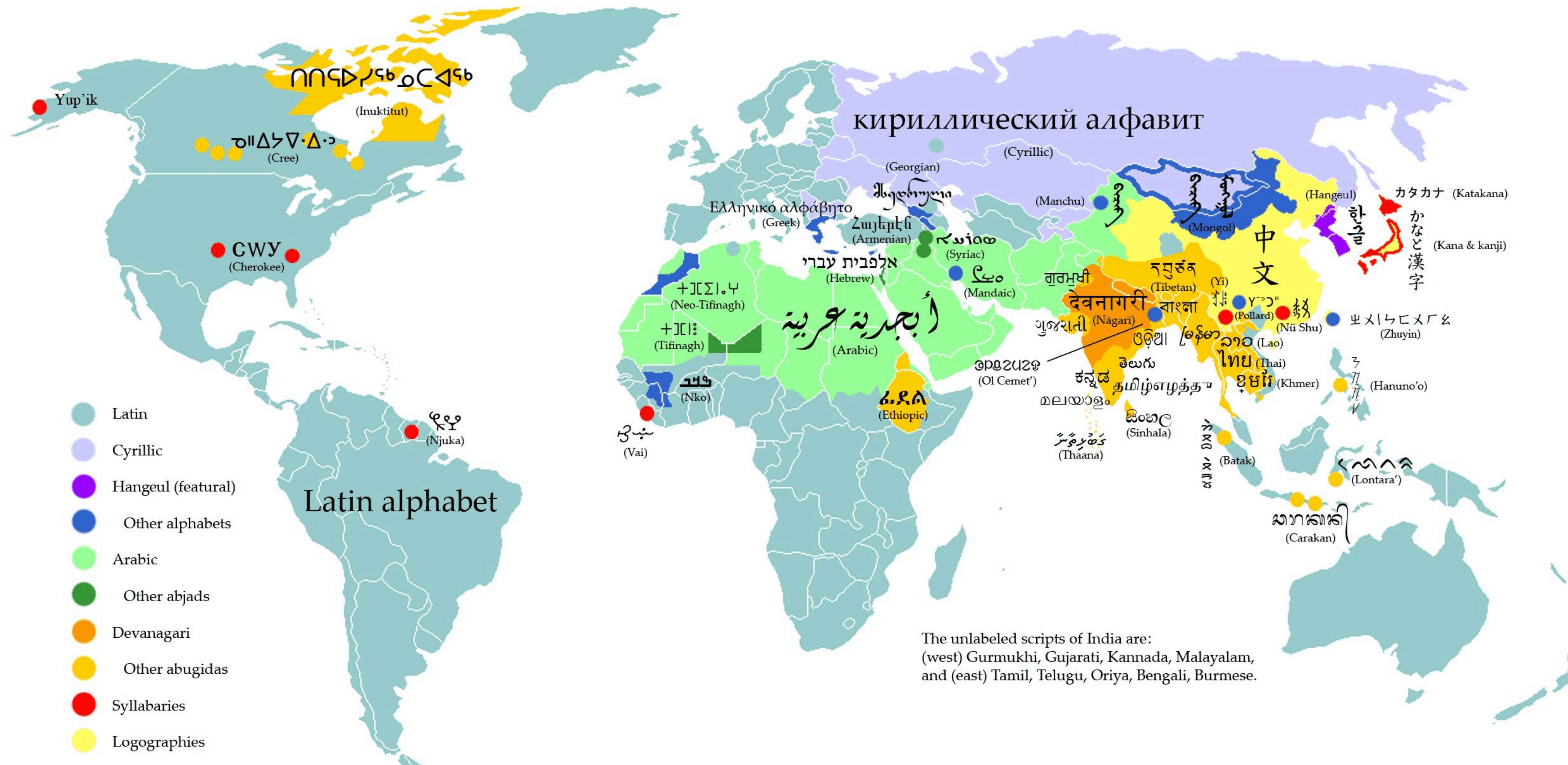
Kumarajiva statue in Xinjiang, China



Xuanzang statue in Xi'an, China

MT helps global information flow

7000 languages in the world



Why automatic Machine Translation?

- Too expensive to hire human translator
 - e.g. touring, shopping, restaurant eating in a foreign country
- Too much effort for human to translate massive text
 - can tolerate imprecise translation
- Need instantaneous translation
 - e.g. in international conference

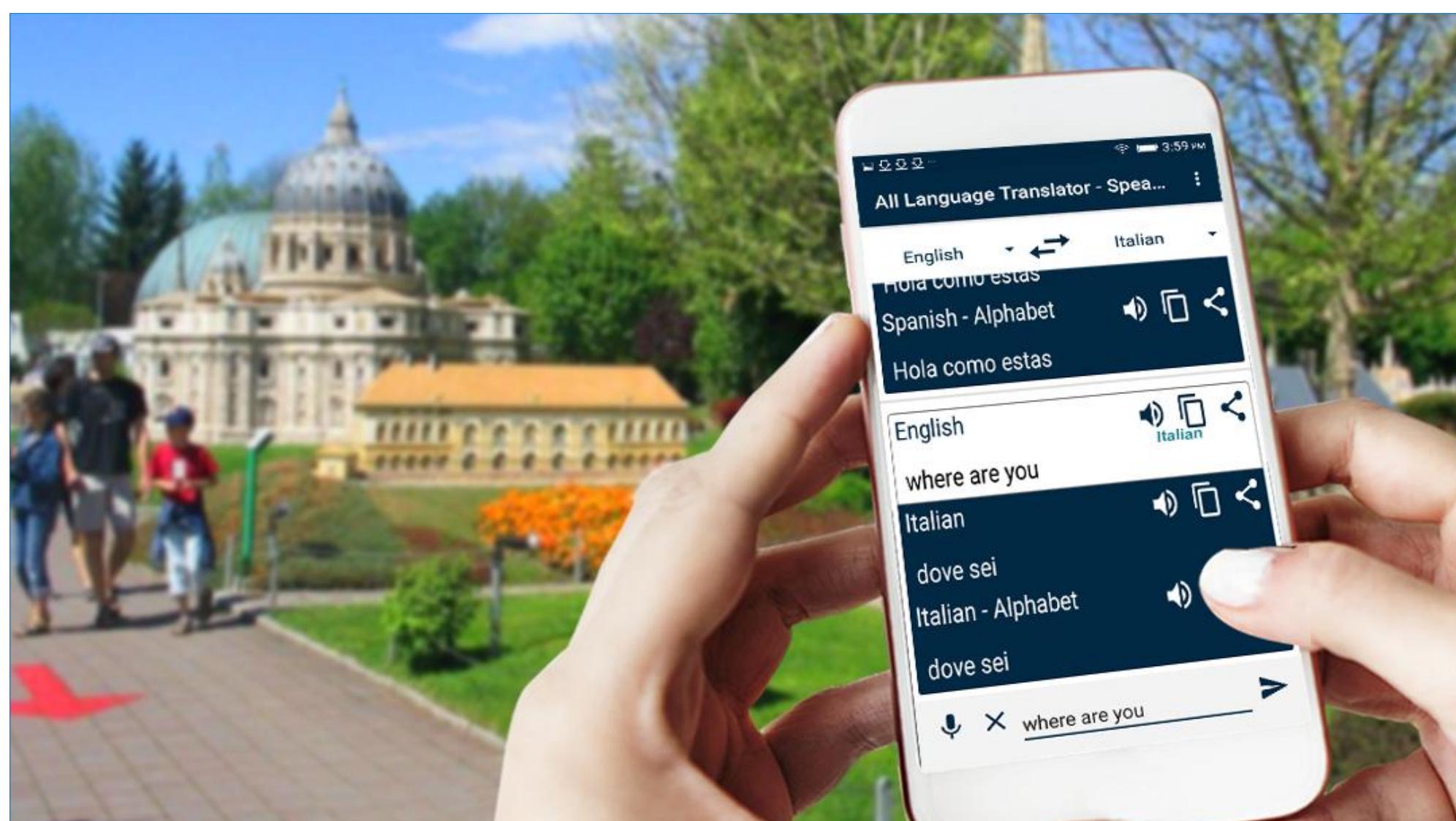
Cross Language Barrier with Machine Translation



Foreign Media



Global Conferences



Tourism



International Trade and e-commerce

Machine Translation has increased international trade by over 10%



<http://pubsonline.informs.org/journal/mnsc>

MANAGEMENT SCIENCE
Vol. 65, No. 12, December 2019, pp. 5449–5460
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

Does Machine Translation Affect International Trade? Evidence from a Large Digital Platform

Erik Brynjolfsson,^a Xiang Hui,^b Meng Liu^b

^a Sloan School of Management, Massachusetts Institute of Technology, Cambridge, Massachusetts 02142; ^b Marketing, Olin School of Business, Washington University in St. Louis, St. Louis, Missouri 63130

Contact: erikb@mit.edu, <http://orcid.org/0000-0002-8031-6990> (EB); hui@wustl.edu, <http://orcid.org/0000-0001-7595-3461> (XH); mengl@wustl.edu, <http://orcid.org/0000-0002-5512-7952> (ML)

Received: April 18, 2019

Revised: April 18, 2019

Accepted: April 18, 2019

Published Online in Articles in Advance:
September 3, 2019

<https://doi.org/10.1287/mnsc.2019.3388>

Copyright: © 2019 INFORMS

Abstract. Artificial intelligence (AI) is surpassing human performance in a growing number of domains. However, there is limited evidence of its economic effects. Using data from a digital platform, we study a key application of AI: machine translation. We find that the introduction of a new machine translation system has significantly increased international trade on this platform, increasing exports by 10.9%. Furthermore, heterogeneous treatment effects are consistent with a substantial reduction in translation costs. Our results provide causal evidence that language barriers significantly hinder trade and that AI has already begun to improve economic efficiency in at least one domain.

History: Accepted by Joshua Gans, business strategy.

Supplemental Material: The online appendix is available at <https://doi.org/10.1287/mnsc.2019.3388>.

Keywords: artificial intelligence • international trade • machine translation • machine learning • digital platforms

Equivalent to
make the
world
smaller than
26%

study on ebay

When you really need Machine Translation

- Rimi Natsukawa live streaming on Tiktok
July, 2021



INA 0 5
CHN 0 10

TOKYO 2020



TOKYO 2020

OMEGA
PIONEER
5-10

OMEGA
PIONEER
5-10

TOKYO 2020



TOKYO 2020

1

TOKYO 2020

1



TOKYO 2020

History of MT

History of Machine Translation

- Warren Weaver: translation as 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."

(1947, in a letter to Norbert Wiener)



1950s-1960s

- 1954: Georgetown-IBM experiment, automatic translation of 60 Russian sentences into English, using lexical rules.
 - Only 6 grammar rules and 250 tokens.
 - W. John Hutchins , Leon Dostert , Paul Garvin
- 1966 ALPAC report
 - We do not have useful machine translation and there is no immediate or predictable prospect of useful machine translation
 - Funding cut for MT in US in the following 20 yrs



Rule-based System

- METEO system for weather forecasts (1976)
 - Used by Environment Canada from 1981 to 2001, to translate between English and French
- Systran (1968)

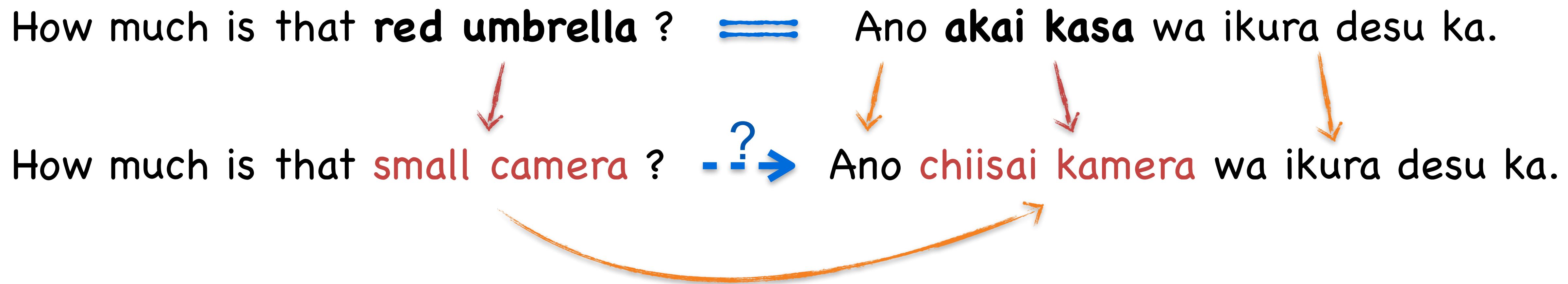
I bought a sweet persimmon in the store

↓ ↓ ↓ ↓ ↓ ↓ ↓

Ich kaufte eine süße Persimone im laden

Example-based Machine Translation

- 1984: Makoto Nagao, A framework of mechanical translation between Japanese and English by analogy principle

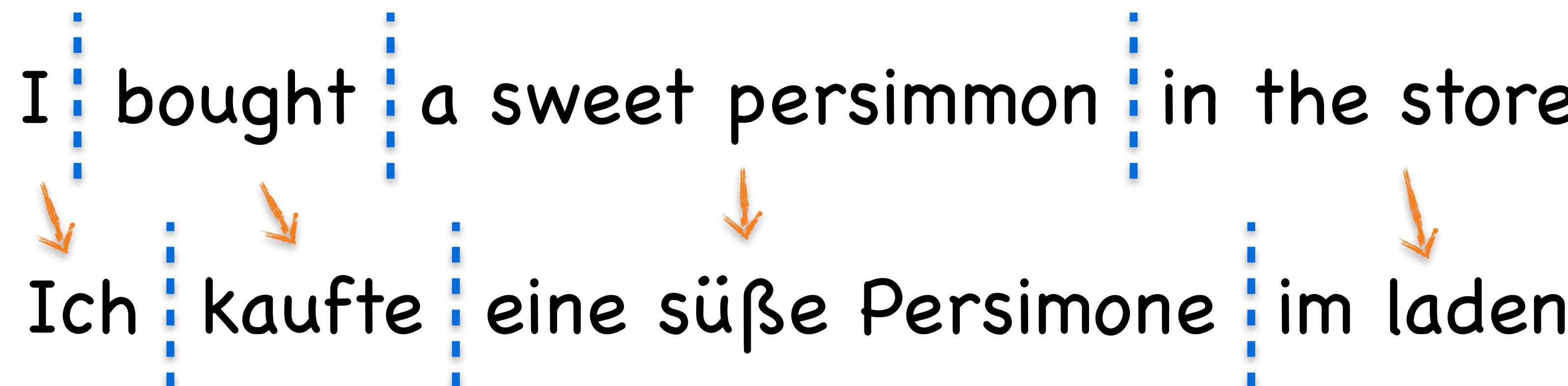


Statistical Machine Translation

- late 1980s-1990s: IBM
- 2000s: phrase-based MT (Moses, Google)
- Training statistical model from parallel corpus

$$\operatorname{argmax} p_{\theta}(y \mid x) = \frac{p(x \mid y)p(y)}{p(x)}$$

$p(x \mid y)$: translation model, $p(y)$: language model



Neural Machine Translation

- Trained in end-to-end fashion (no intermediate separate training)
- 2014: Sequence to sequence learning with neural networks
 - Define LSTM encoder-decoder framework
- 2015: Neural Machine Translation by Jointly Learning to Align and Translate
 - Define attention mechanism between encoder-decoder
- 2016: Google translate deploys NMT
- 2017: Attention is all you need
 - Replace LSTM with multihead attention layers (Transformer)
- Almost all major production MT systems use NMT now

MT Products

- Google translate: 109 languages, separate app, support text/document translation, image translation, and speech translation
- Microsoft translate: 87 languages for text
- Baidu translate: 200+ languages
- ByteDance VolcTrans: 55 languages
- DeepL: good at European languages
- Youdao Translate: integrated with its own dictionary app
- Tencent Translate:
- NiuTrans: specialized in Chinese to many languages

MT Products for users/clients

Web translate tool, Translation function on Youtube/Tiktok/Twitter/ Facebook	consumers/users who do not know the source language	could tolerate imprecision	convenient, free/low-cost
Computer Aided Translation tools	content creator, translators, knowing both languages	need high precision	productivity and efficiency, additional functionality like translation memory, glossary
translation API e.g. Amazon translation	business client	cost/effective	robust api, easy to integrate and maintain
private MT deployment e.g. NiuTrans	business client		domain-specific models, tailored to special needs
Special MT hardware, e.g. translation pen Simultaneous translation earphone	consumers for targeted scenario		

MT is not just about Model

- User-oriented Product
 - What are real users' needs?
 - Observe how the users are using our product, e.g. how translators are using CAT tools
- Data-oriented
 - Look at the cases translated by systems
 - Not just automatic metric
- System-oriented
 - Building high-performance, reliable, easy-to-maintain system

MT Cases and Challenges

Examples

周四经济数据面，美国劳工部报告称，截至8月28日当周美国首次申请失业救济人数为34万，降至2020年美国新冠疫情危机爆发以来的最低点。市场预计该数字为34.5万。

Google Translation (2021.9.1)

On Thursday's economic data, the U.S. Department of Labor reported that as of August 28, the number of people applying for unemployment benefits for the first time was 340,000, which dropped to the lowest point since the outbreak of the new crown crisis in the United States in 2020. The market expects the number to be 345,000.

VolcTrans (2021.9.1)

On Thursday's economic data, the U.S. Labor Department reported that the number of first-time jobless claims in the United States for the week ending August 28 was 340 thousand, falling to the lowest level since the COVID-19 Epidemic broke out in the United States in 2020. The market expects the number to be 345 thousand.

Examples

周四经济数据面，美国劳工部报告称，截至8月28日当周美国首次申请失业救济人数为34万，降至2020年美国新冠疫情危机爆发以来的最低点。市场预计该数字为34.5万。

Bing Translation (2021.9.1)

On Thursday, the *Labor Department reported that 340,000 people applied for *unemployment benefits for the week ended Aug. 28, the lowest level since the *crisis began in 2020. The market expects the figure to be 345,000.

DeepL (2021.9.1)

On Thursday's economic data front, the U.S. Labor Department reported that the number of first-time U.S. jobless claims for the week ended Aug. 28 was 340,000, falling to the lowest point since the outbreak of the new U.S. crown epidemic crisis in 2020. The market expected the figure to be 345,000.

Examples

周四美股成交额冠军苹果(153.65, 1.14, 0.75%)公司收高0.75%，报153.65美元，创历史新高，收盘成交108.9亿美元，市值逼近2.54万亿美元。

Google Translation (2021.9.1)

Apple (153.65, 1.14, 0.75%), the champion of U.S. stock market turnover on Thursday, closed 0.75% higher at US\$153.65, a record closing high, with a turnover of US\$10.89 billion and a market value of approximately US\$2.54 trillion.

VolcTrans (2021.9.1)

U.S. stock turnover champion Apple (153.65, 1.14, 0.75%) closed up 0.75% at \$153.65 on Thursday, a record closing high of \$10.89 billion and a market value approaching \$2.54 trillion.

Examples

周四美股成交额冠军苹果(153.65, 1.14, 0.75%)公司收高0.75%，报153.65美元，创历史新高，收盘成交108.9亿美元，市值逼近2.54万亿美元。

Bing Translation (2021.9.1)

U.S. stock market champion Apple Inc (153.65, 1.14, 0.75 percent) closed up 0.75 percent at \$153.65 on Thursday, a record closing high of \$10.89 billion, giving it a market capitalization of nearly \$2.54 trillion.

DeepL (2021.9.1)

Thursday's U.S. stock turnover leader Apple (153.65, 1.14, 0.75%) closed 0.75% higher at \$153.65, an all-time closing high, with \$10.89 billion traded and a market cap approaching \$2.54 trillion.

Why MT is challenge

- Polysemy

He deposited money in a **bank** account
with a high **interest** rate.

Sitting on the **bank** of the Mississippi, a
passing ship piqued his **interest**.

- New entity names
 - COVID-19
- Complex structure
- Ellipsis (i.e. omission)

他的爷爷和奶奶没见过他的姥姥和姥爷。

Google Translate: His grandpa and grandma have
never met his grandma and grandpa.

Correct: His father's parents never met his mother's.

- Acronym and incorrect word segmentation

一些立陶宛人士表示，中立关系恶化，影响最大的当属立陶宛的出口企业。

Google Translate: Some Lithuanians said that the deterioration of Sino-Lithuanian relations has affected Lithuanian export companies the most.

Bing Translate: Some Lithuanians say the deterioration in neutral relations has affected Lithuania's exporters the most.

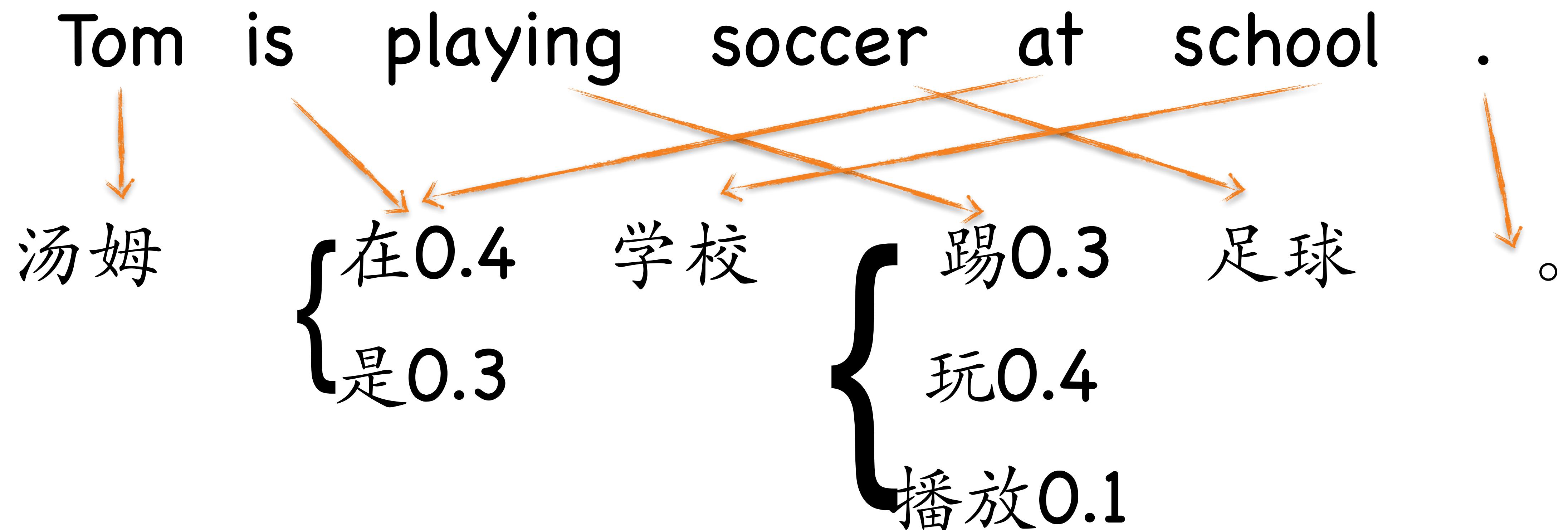
Basics

A Statistical Model for MT

$$\operatorname{argmax} p_{\theta}(y|x) \propto p(x|y)p(y)$$

translation probability

language model



(Statistical) Language Model

- Estimating $p(y)$ in a target language

$$p(y) = p(y_1, y_2, \dots, y_n) = \prod_{t=1}^n p(y_t | y_1, \dots, y_{t-1})$$

- Many ways to estimate $p(y_t | y_1, \dots, y_{t-1})$
- Direct estimate n-gram statistics $p(y_t | y_{t-k}, \dots, y_{t-1})$
- e.g. Uni-gram $p(\text{"hello"})$, $p(\text{"world"})$,
- e.g. bi-gram $p(\text{"world"} | \text{"hello"})$
- What if the frequency of n-gram is too small? Smoothing.

Probability Basics

- Marginal probability: $p(B) = \sum_A P(A, B)$

- Conditional probability: e.g. bigram prob.

$$p(B | A) = \frac{P(A, B)}{P(A)}$$

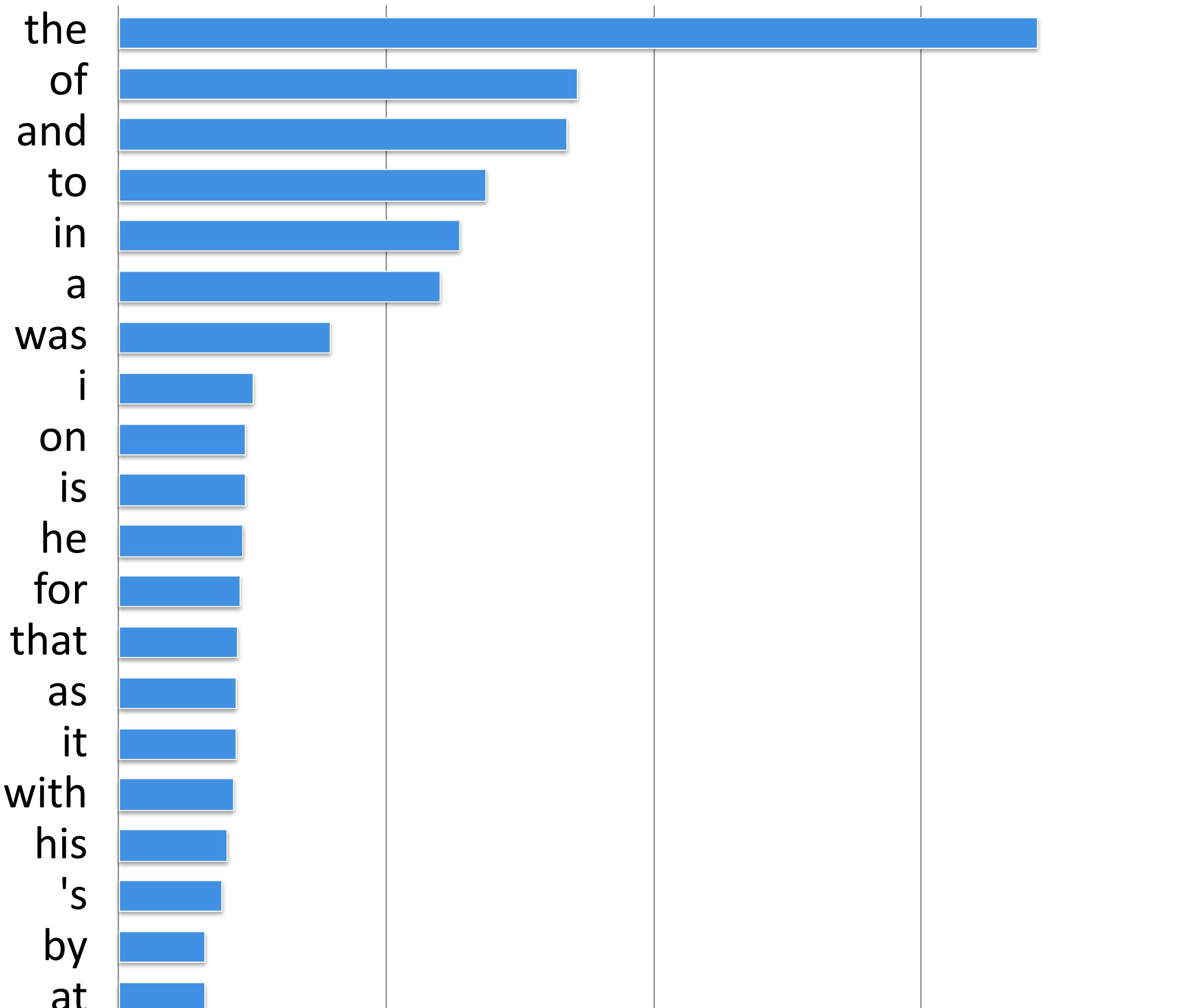
$$p(\text{gas} | \text{reduce}) = \frac{\text{freq}(\text{reduce}, \text{gas})}{\text{freq}(\text{reduce})}$$

- Bayes rule: $p(B | A) = \frac{p(A | B)p(B)}{p(A)}$

– the most important formula in AI!

Word and Bigram

Statistics from English Wikipedia and books



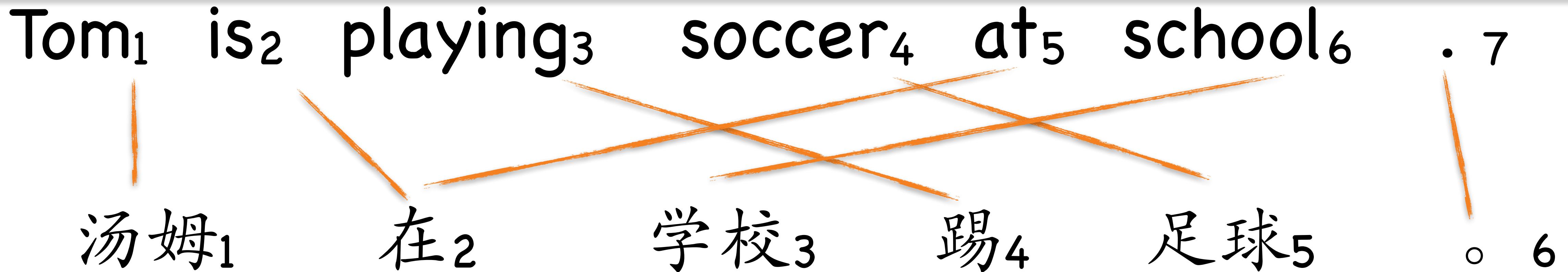
cond. prob. $p(x_2|x_1)$

	first	united	the	a	be
the	0.014	0.006			
of			0.283	0.030	
would					0.191
with			0.187	0.122	

Challenge of n-gram LM

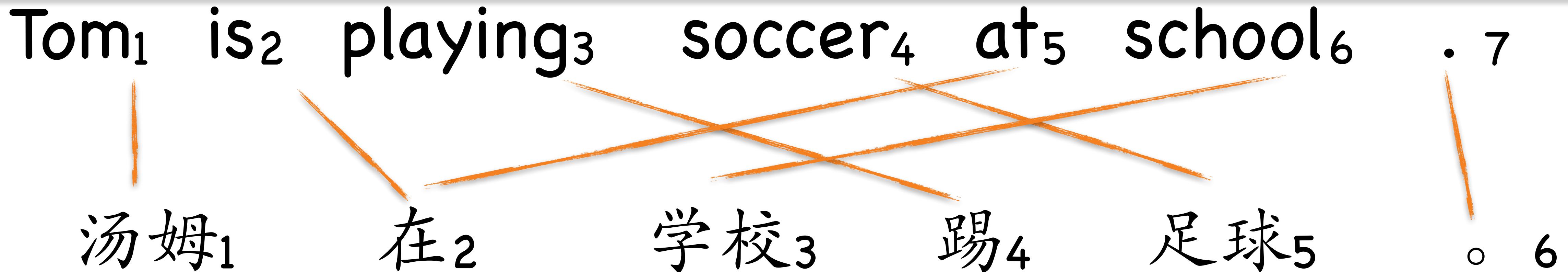
- Vocabulary: V
- n-gram needs a probability table of size V^n
- Common V size $30k \sim 100k$
- Hard to estimate and hard to generalize
- Solution: Parameterization with generative model
 - $p(y_t | y_1, \dots, y_{t-1}; \theta) = f_\theta(y_1, \dots, y_{t-1})$
 - f can be a carefully designed and computationally tractable function, e.g. a neural network (later lectures).

Translation Probability



- The translation prob. $p(x | y) = \sum_a p(x, a | y)$
- a is alignment from source (English) to target (Chinese): each word in English sentence is mapped to one word (or null) in Chinese, denoted by index.
- a={1, 2, 4, 5, 2, 3, 6}

Translation Probability



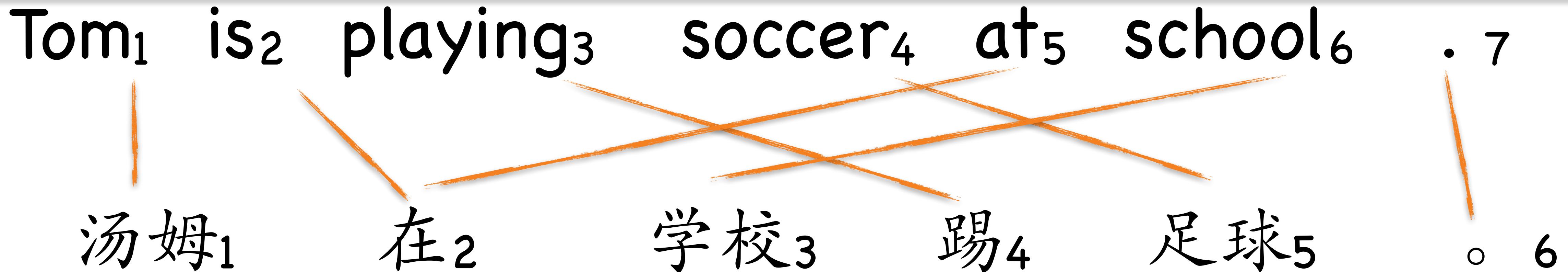
- The translation prob. $p(x|y) = \sum_a f(x, a | y)$

- m: source sent. length, l: target sent. length

$$p(x, a | y) = p(m | y) \prod_{j=1}^m p(a_j | a_1 \dots j-1, x_1 \dots j-1, m, y) p(x_j | a_1 \dots j, x_1 \dots j-1, m, y)$$

- IBM model 1: $p(x, a | y) = \frac{\epsilon}{(l+1)^m} \prod_{j=1}^m t(x_j | y_{a_j})$

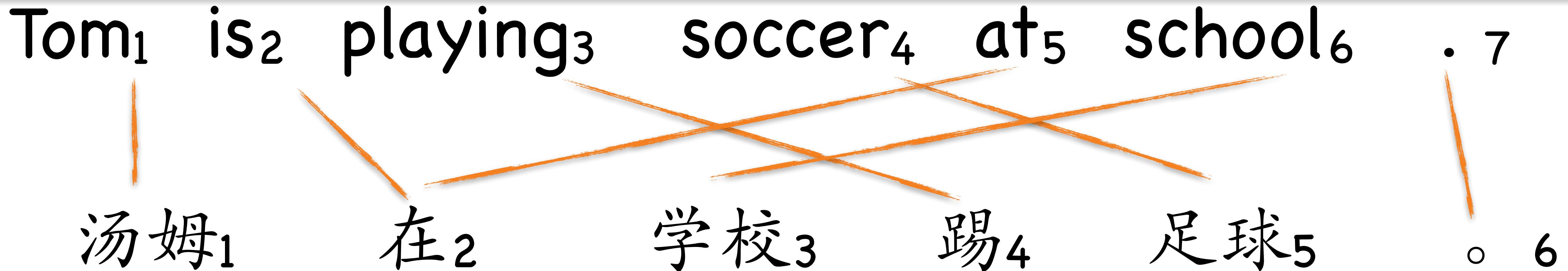
Translation Probability



- The translation prob. $p(x|y) = \sum_a f(x, a|y)$
- IBM model 1:

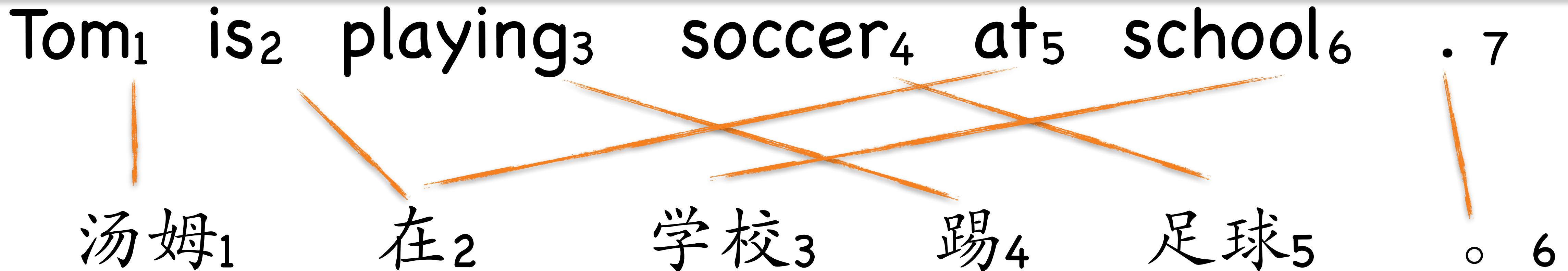
$$p(x|y) = \frac{\epsilon}{(l+1)^m} \sum_{a_1=0}^l \cdots \sum_{a_l=0}^m \prod_{j=1}^m t(x_j | y_{a_j})$$

Learning Parameters



- The translation prob. $p(x | y) = \sum_a f(x, a | y)$
- Model parameter $\theta = \{t_{ji}\} \leftarrow \operatorname{argmax} \sum_n \log p(x_n | y_n)$
- Expectation-maximization algorithm
 - taking derivative and equating to zero

Learning Parameters



- Expectation-maximization algorithm
 - taking derivative and equating to zero

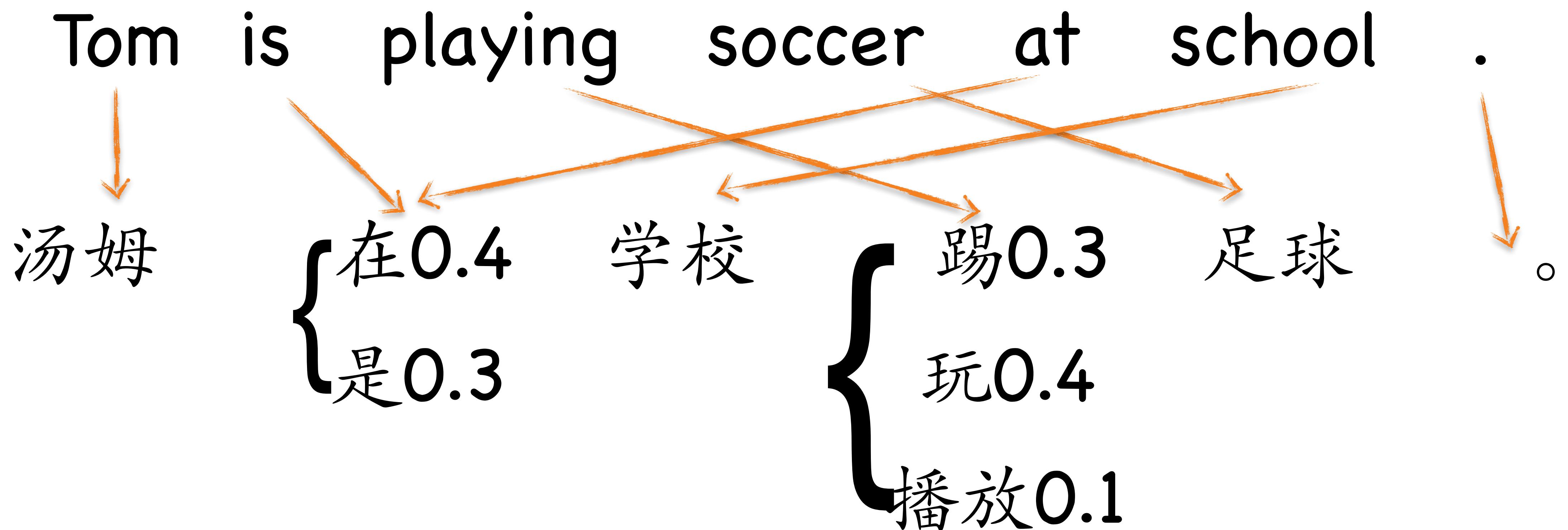
$$\theta = \{t_{wv}\} = \lambda_v^{-1} \sum_a p(x, a | y; \theta) \underbrace{\sum_{j=1}^m \delta(w, x_j) \delta(v, y_{aj})}_{\text{number of times a source word } w \text{ connects a target word } v \text{ in sentence pair } (x, y)}$$

number of times a source word w connects a target word v in sentence pair (x, y)

The full SMT model

$$\operatorname{argmax} p_{\theta}(y|x) \propto p(x|y)p(y)$$

translation probability language model



More to Consider

- The first model IBM Model 1 is over simplified with a lot of independence assumptions

IBM Model 1	Lexical model
IBM Model 2	global alignment model, alignment is dependent on position
IBM Model 3	adding fertility model
IBM Model 4	relative reordering model
IBM Model 5	fixes deficiency

Where to find papers?

- General NLP conferences:
 - ACL: Annual Meeting of the Association for Computational Linguistics
 - EMNLP: Conference on Empirical Methods in Natural Language Processing
 - NAACL: North American Chapter of the Association for Computational Linguistics
 - AACL, EACL, COLING, CoNLL, NLPCC, CCL
- General ML/AI conference:
 - ICML, NeurIPS, ICLR, AAAI, IJCAI also have MT papers
- MT Conference:
 - WMT: Conference on Machine Translation, used to be Workshop on Statistical Machine Translation
 - IWSLT: International Conference on Spoken Language Translation
- NLP journals
 - Transaction on ACL
 - Computational Linguistics

Data Resources

- (Text) Machine Translation:
 - News Translation (general domain): <http://statmt.org/wmt21/translation-task.html>
 - includes data from many sources:
 - Europarl
 - UN Parallel Corpus
 - CommonCrawl: both parallel aligned and raw data
 - OPUS: <https://opus.nlpl.eu/index.php>
- Speech Translation:
 - MuST-C: <https://ict.fbk.eu/must-c/>
 - CoVoST: <https://github.com/facebookresearch/covost>
 - LibriSpeech
- Tatoeba: collections of translations, <https://tatoeba.org/en>
- Wikipedia: raw corpus

Software and Library

- Pytorch: <https://pytorch.org/>
- Tensorflow: <https://www.tensorflow.org/>
- JAX: <https://github.com/google/jax>
- NeurST: <https://github.com/bytedance/neurst>
- FairSeq: <https://github.com/pytorch/fairseq>
- Huggingface: <https://github.com/huggingface/transformers>
- LightSeq: <https://github.com/bytedance/lightseq>

Language in 10mins

- [https://sites.cs.ucsb.edu/~lilei/course/dl4mt21fa/
LanguagePresentation.html](https://sites.cs.ucsb.edu/~lilei/course/dl4mt21fa/LanguagePresentation.html)
- work in group (two person, different from project)
- Survey a language (be diverse, favor low-resource languages)
- give 10 mins presentation in-class
- Linguistic and other useful facts
- Machine Translation system (if available) and performance.
- Pick a slot to present (today): <https://tinyurl.com/4m8yjkuv>
- Turn-in your short write-up (up to 4 pages)
- [https://en.wikipedia.org/wiki/
List_of_languages_by_number_of_native_speakers](https://en.wikipedia.org/wiki/List_of_languages_by_number_of_native_speakers)

MT Blog

- Write a popular science/MIT technology review article about one paper in MT
- Read one paper in detail (not in group, but can discuss with other students)
 - choose from the suggested list or your own choice, confirm with Instructor
- Try to Reproduce results
 - no need to re-train
 - but need to use their published model to inference on same or extra data
 - Case study
- Indicate whether ok to put public
- Deadline: 6 sessions from LangPresentation slot if before 11/10 (but no later than 11/22), 11/1 if LangPresentation slot after 11/15.

MT Blog

- Writing suggestion
 - In Markdown (with math support), or HTML (w/ javascript, no php)
 - more than 1/4 of content (the problem, challenge, intuition etc) should be understood by high school students (layman's term, e.g. if you present to your mother/father/grandparents)
 - about 1/2 of content understood by college students
 - no more than 1/4 of content understood by MT researchers
 - Use visualization, figures, tables, and show-case examples
 - Interactive (e.g. via js) could be helpful as well

MT Blog

- Writing Template:
 - VuePress template: <https://github.com/lileicc/blog>
 - ▶ You may create and edit a new markdown file under blogs/ directory
 - ICLR blog template: <https://iclr-blog-track.github.io/submitting/>
- Example:
 - <https://lileicc.github.io/blog/blogs/mt/2020/mrasp>
 - <https://towardsdatascience.com/byte-pair-encoding-subword-based-tokenization-algorithm-77828a70bee0>
 - https://lena-voita.github.io/posts/nmt_training_through_smt_lens.html

MT Blog

- (Optional) ICLR 2022 Blog Track
 - You may consider submit to ICLR 2022 Blog Track
 - <https://iclr.cc/Conferences/2022/CallForBlogPosts>
 - But be sure to keep very high quality
 - You may ask for recommendation whether appropriate for submission from instructor
 - Avoidance of conflicts of interest
 - Author of the papers
 - Same institution

Project

- Work in 2-3 person group.
- Conduct a research project in machine translation.
- Reasonable size that can be completed within quarter.
- Discuss with Instructor.
- Proposal due: due 10/18
 - What is your plan topic, what is current status of the problem, what is the challenge, what are your potential methods, which data do you plan to evaluate on
- Mid-term report: 11/10
 - Progress so far, what are the problems encountered, how do you plan to solve, any risks
- Final Project Report: 12/1
 - A full report with problem statement, motivation, related work, proposed method, results, what are impacts
 - Poster presentation on 12/1 (not necessary same time, to be arranged in open space)

Project inspiring ideas

- Develop a working MT system for some new (no high-quality available MT) and low-resource languages (e.g. Spanish-to-Tamil), explore and solve challenges along the way
- Improving methods to better utilize monolingual data
- Extending and improving Vocabulary and Tokenization for NMT
- Improving evaluation quality and efficiency, certain human-assisting tools for evaluation, conduct study.
- Computer-assisted and interactive translation methods
- MT for multimodal data, e.g. video translation, speech translation
- Integrating domain knowledge into MT system
- Novel hardware-based MT system, e.g. Compress MT model to very small size and build a system (with inference but not training) on mobile phones, or extending existing CUDA library (e.g. LightSeq) to support more complex models.
- Application of MT technology to your own project

Computing Resource

- Cloud computing: thanks to AWS
 - Each student gets \$300 credit on AWS cloud for this course (about 400 hrs of training on a gpu machine)
 - <https://awsacademy.instructure.com/courses/7966/>
 - You must agree to the terms <https://aws.amazon.com/legal/learner-terms-conditions/> to be able to use.
- Translation platform support from VolcTrans (still in negotiation)



Reading

- Peter F. Brown and Stephen A. Della-Pietra and Vincent J. Della-Pietra and Robert L. Mercer (1993): The Mathematics of Statistical Machine Translation: Parameter Estimation, Computational Linguistics
- (or) Chap 4 in PK SMT book.
- Warren Weaver, Translation, 1949.