Machine Translation and Sequence-to-sequence Models

http://phontron.com/class/mtandseq2seq2019/

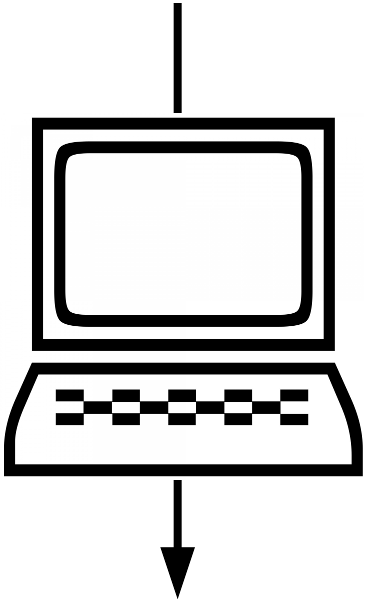
Graham Neubig and Antonios Anastasopoulos

Carnegie Mellon University

CS 11-731

What is Machine Translation?

kare wa ringo wo tabeta .



He ate an apple .

What are Sequence-to-sequence Models?

Sequence-to-sequence Models

Machine translation:

kare wa ringo wo tabeta → he ate an apple

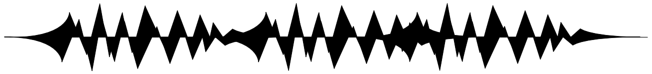
Tagging:

he ate an apple → PRN VBD DET PP

Dialog:

he ate an apple → good, he needs to slim down

Speech Recognition

 → he ate an apple

And just about anything...:

1010000111101 → 00011010001101

Why MT as a Representative?

Useful!

KUSHINIKIZA! Google Translate SAVES BABY in Irish roadside birth

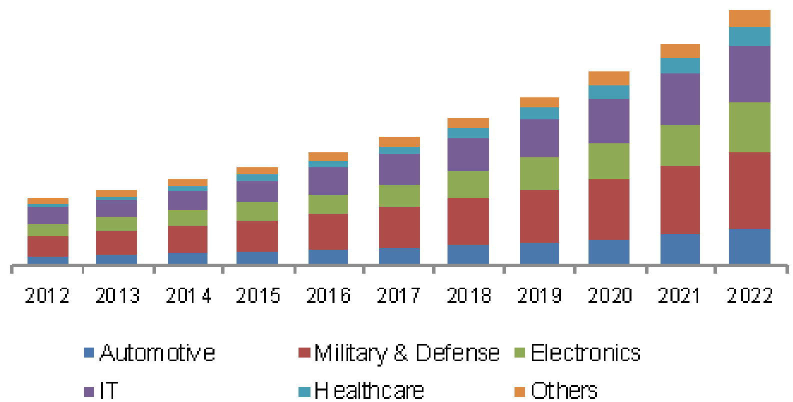
Do no evil? We literally save lives now

13 Feb 2015 at 12:01, John Leyden 8

Quick-thinking Irish paramedics turned to Google Translate to communicate with a pregnant woman who spoke Swahili, allowing her to safely give birth.

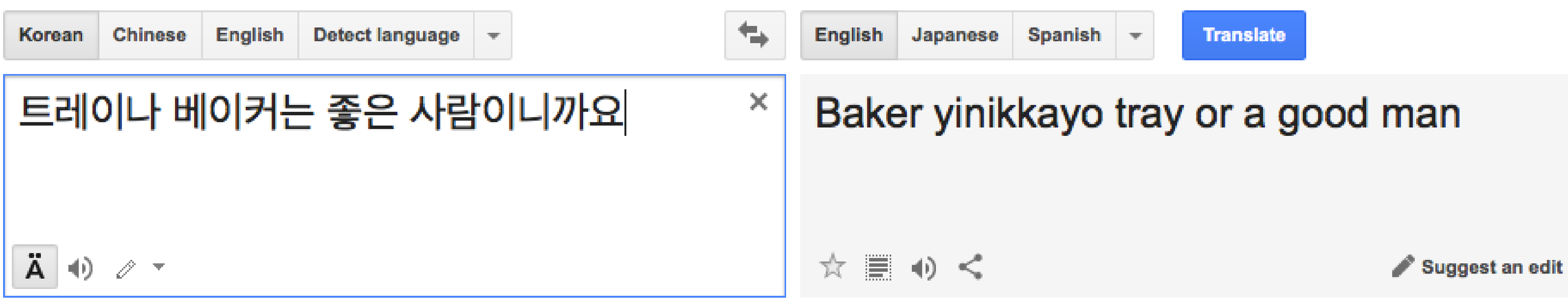
Source: The Register

Global MT Market Expected To Reach $983.3 Million by 2022



Source: Grand View Research

Imperfect...



MT and Machine Learning

Big Data! Billions of words for major languages
… but little for others

Well-defined, Difficult Problem!

Use for algorithms, math, etc.

Algorithms Widely Applicable!

MT and Linguistics

트레이나 베이커는 좋은 사람이니까요

Baker yinikkayo tray or a good man

Trina Baker is a good person

Morphology! 이니까요 is a variant of 이다 (to be)

Syntax! should keep subject together

Semantics! "Trina" is probably not a man...

... and so much more!

Class Organization

Class Format

- **Before class:**
  - Read the assigned material
  - Ask questions via web (piazza/email)
- **In class:**
  - Take a small quiz about material
  - Discussion, questions, elaboration
  - In some cases, code-walk

Assignments

- **Assignment 1:** Create a neural sequence-to-sequence modeling system. Turn in code to run it, and write a report.
- **Assignment 2:** Create a system for a challenge task, to be decided in class (maybe low-resource translation).
- **Final project:** Come up with an interesting new idea and test it.

Assignment Instructions

- Assignment 1: individual
  Assignment 2, Project: work in groups of 2-3.
- Use a **shared git repository and commit the code that you write**, and in reports note who did what part of the project.
- All implementations must be basically your own, although you can use small code snippets if you cite them.
- We recommend implementing in Python (PyTorch, DyNet)

Class Grading

- **Short quizzes:** 20%
- **Assignment 1:** 20%
- **Assignment 2:** 20%
- **Final Project:** 40%

Class Plan

1. Introduction (Today): 1 class
2. Language Models: 2 classes
3. Neural MT: 3 classes
4. Evaluation/Analysis: 1 class
5. Linguistically Informed Models, Multilingual Learning: 4 classes
6. Advanced Learning Methods: 2 classes
7. Applications: 4 classes
8. Symbolic MT: 3 classes
9. Advanced Topics: 3 classes
10. Final Project Presentations: 2 classes

Guest Lectures

- **Chenhui Chu** (Osaka University, 9/19): Adaptation Methods
- **Niki Parmar** (Google Brain, 10/15): Advances in Tansformers
- **Marcin Junczys-Dowmunt** (Microsoft Research, 10/17): Building State-of-the-art MT Systems

Models for Machine Translation

Machine Learning for Machine Translation

F = *kare wa ringo wo tabeta .*

↓

E = He ate an apple .

Probability model: $P(E|F;\Theta)$

↑

Parameters

Problems in MT

- **Modeling:** How do we define $P(E|F;\Theta)$?
- **Learning:** How do we learn Θ?
- **Search:** Given F, how do we find the highest scoring translation?

$$E' = \operatorname{argmax}_E P(E|F;\Theta)$$

- **Evaluation:** Given E' and a human reference E, how do we determine how good E' is?

Neural MT Basics

Language Models 1: n-gram Language Models

Given multiple candidates, which is most likely as an English sentence?

E_1 = he ate an apple

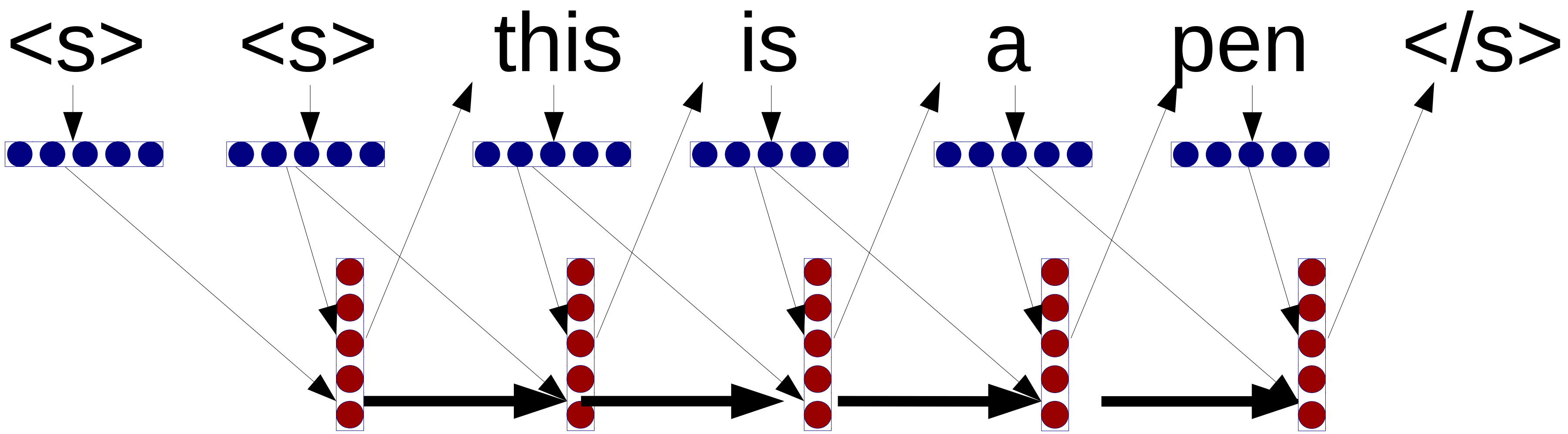
E_2 = he ate an apples

E_3 = he insulted an apple

E_4 = preliminary orange orange

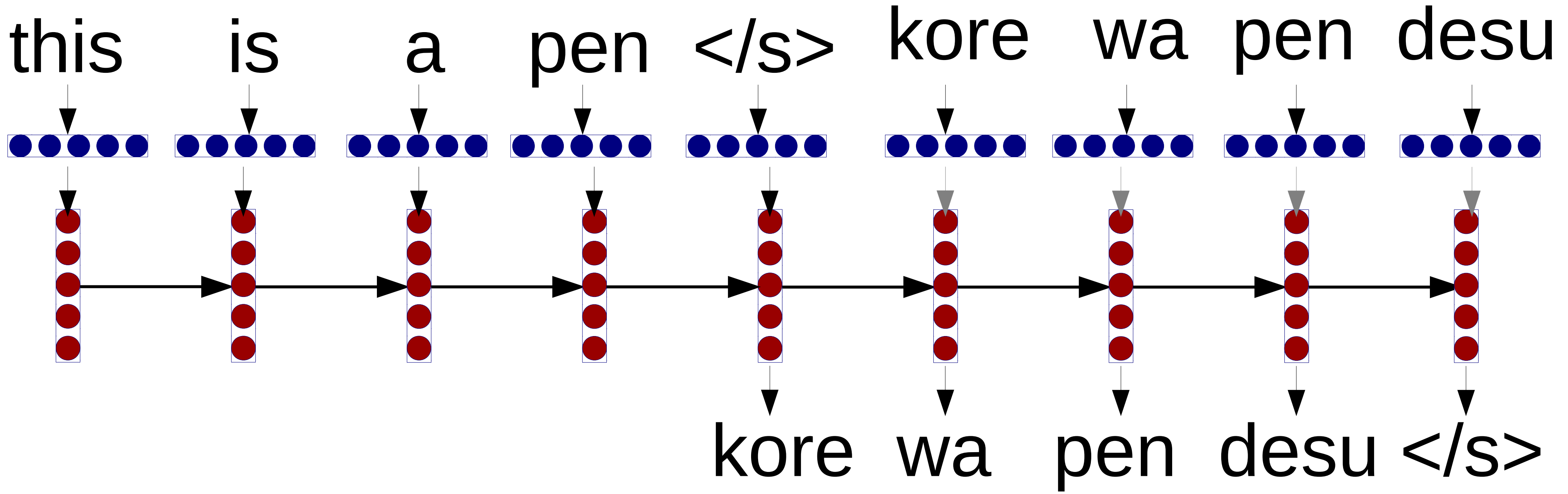
- Definition of language modeling
- Count-based n-gram language models
- Evaluating language models
- **Code:** n-gram language model

Language Models 2: Recurrent LMs



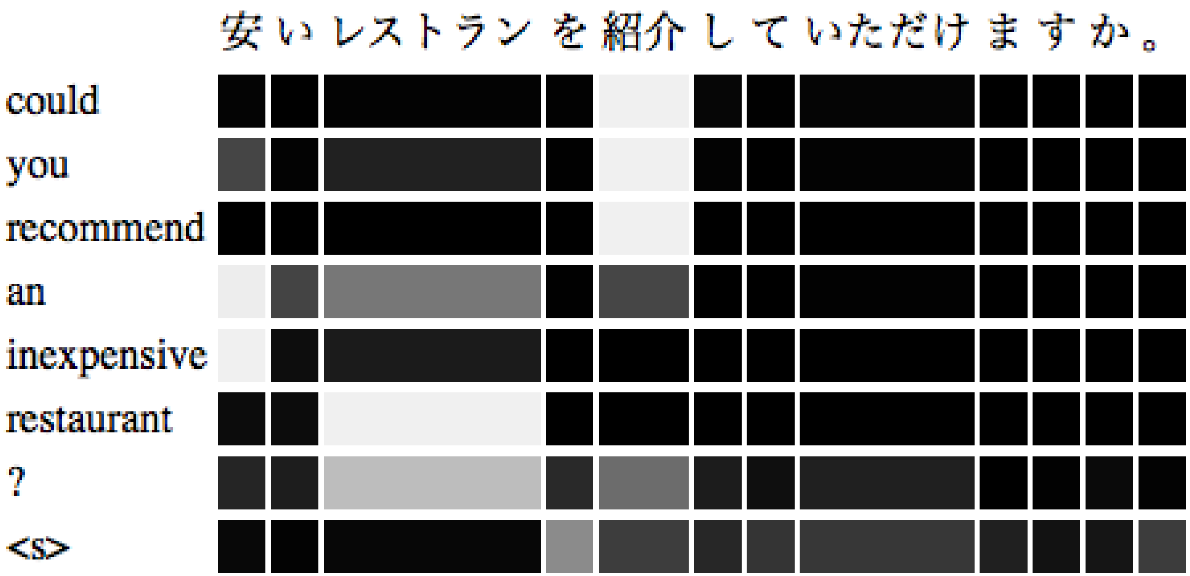
- Neural networks
- Recurrent networks, LSTMs/GRUs
- Training tricks
- **Code:** RNN-based LM

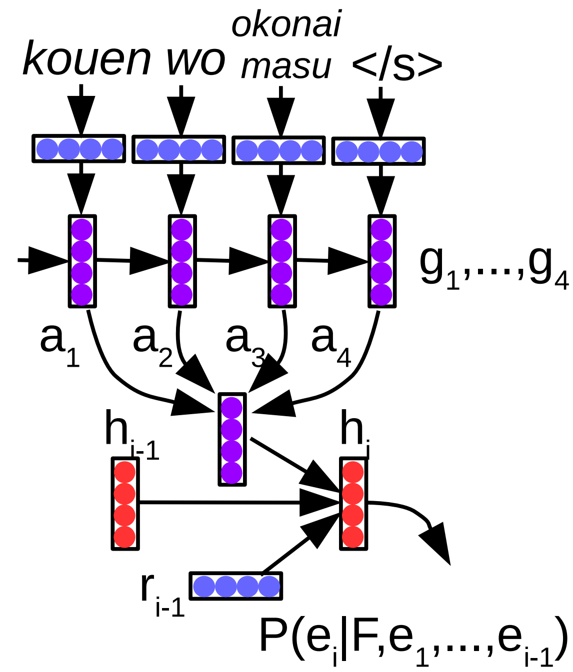
Neural MT 1: Encoder-decoder Models



- Encoder-decoder Models
- Searching for hypotheses
- Mini-batched training
- **Code:** Encoder-decoder model

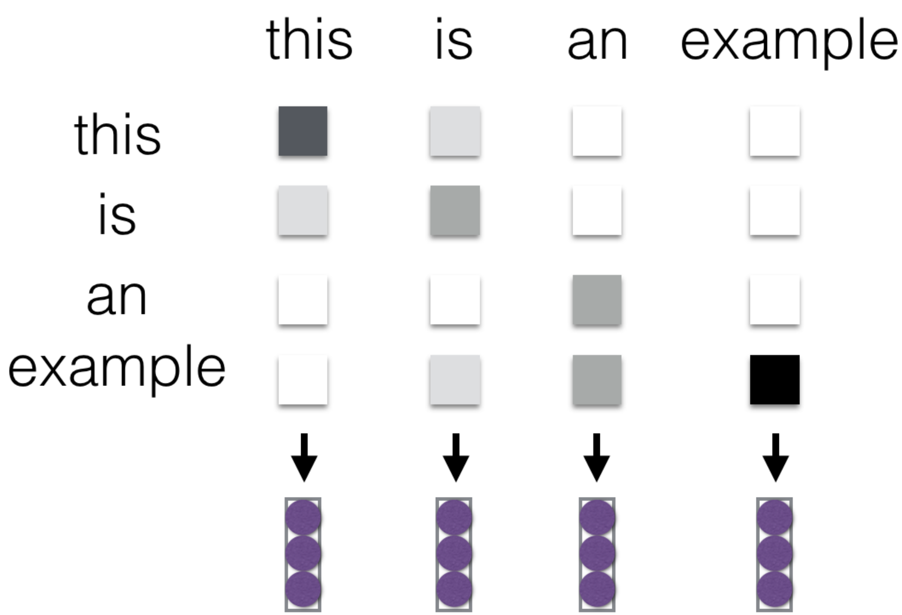
Neural MT 2: Attentional Models





- Attention in its various varieties
- Unknown word replacement, copying
- Attention improvements, coverage models
- **Code:** Attentional model

Neural MT 3: Self-attention, CNNs



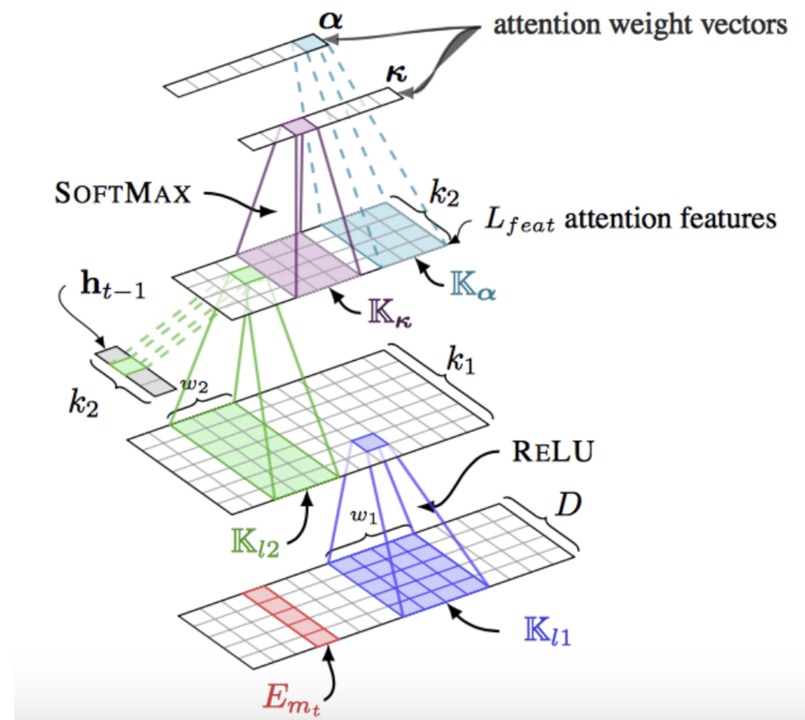


Image: Allamanis et al. (2017)

- Self attention
- Convolutional neural networks
- A case study, the transformer
- **Code:** Self-attentional model

Data and Evaluation

Evaluation

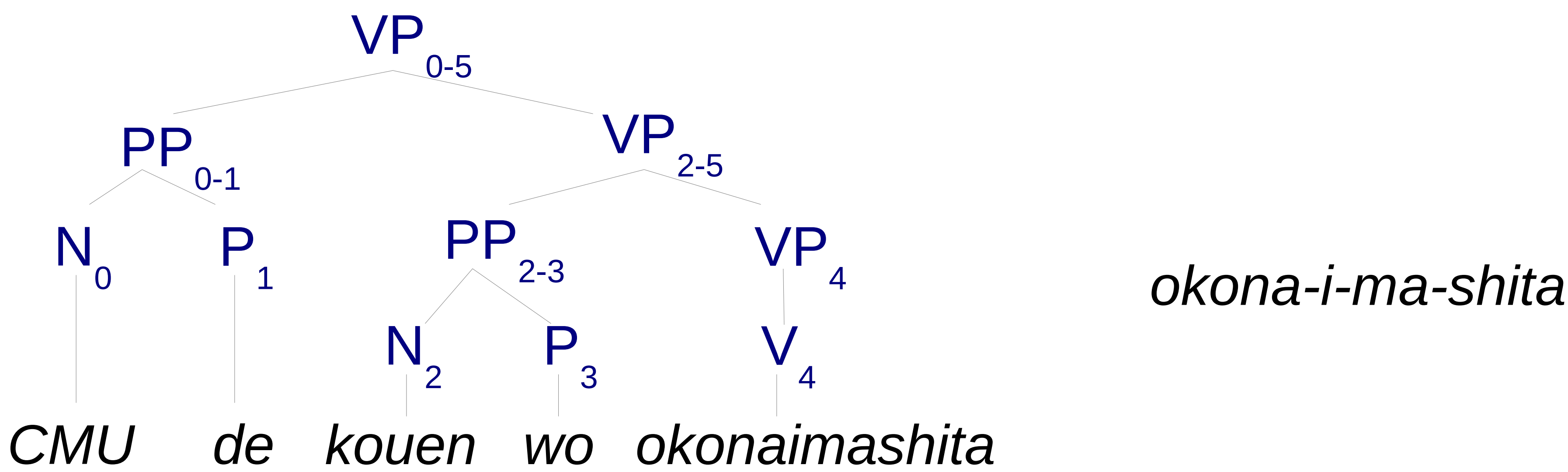
taro ga hanako wo otozureta

	Taro visited Hanako	the Taro visited the Hanako	Hanako visited Taro
Adequate?	○	○	X
Fluent?	○	X	○
Better?	B, C	C	

- Human evaluation
- Automatic evaluation
- Significance tests and meta-evaluation
- **Code:** BLEU, and correlation

Language and Linguistic Models

Syntax/Morphology



- Phrase structure and dependency grammar
- Morphology basics
- Universal dependencies
- **Code:** Running universal dependency parser

Subword Models

reconstructed

↓

re+ construct+ ed

- Character models
- Subword models
- Morphology models
- **Code:** Implement subword splitting

Languages of the World

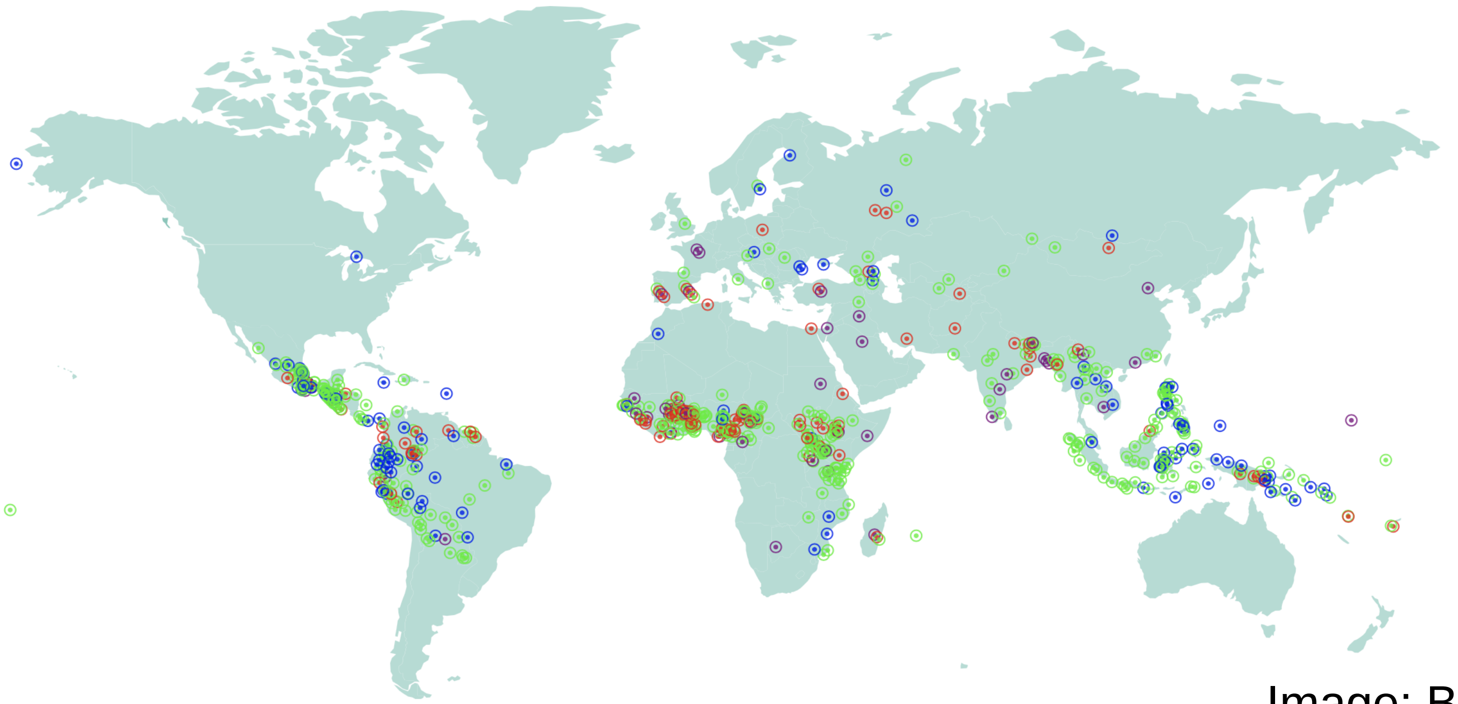


Image: Black (2019)

- Introduction to languages around the world
- Linguistic databases, data sources
- **Code:** Play with large multilingual databases

Multi-lingual Learning

hello

こんにちは

hola

- Learning for multiple languages
- Cross-lingual sharing of syntax or lexicon
- **Code:** Implement a multi-lingual neural system

Advanced Training Techniques

Parameter Optimization

	LM	TM	RM	Highest ○
○ Taro visited Hanako	0.2*-4	0.3*-3	0.5*-1	-2.2
X the Taro visited the Hanako	0.2*-5	0.3*-4	0.5*-1	-2.7
X Hanako visited Taro	0.2*-2	0.3*-3	0.5*-2	-2.3

- Loss functions
- Deciding the hypothesis space
- Optimization criteria
- **Code:** Minimum risk training of MT

Semi- and Unsupervised Learning

- Semi-supervised learning
- Back-translation
- Unsupervised learning
- **Code:** Semi-supervised learning of MT

Application Examples

Applications 1: Paraphrasing, Style and Attribute Transfer

Source	Speaker	Input	Output
Romeo & Juliet	Benvolio	He killed your relative, brave Mercutio, and then young Romeo killed him.	he slew thy kinsman , brave mercutio , and then young romeo kill him .
Romeo & Juliet	Romeo	I can read my own fortune in my misery.	i can read mine own fortune in my woes .

Image: Xu et al. (2011)

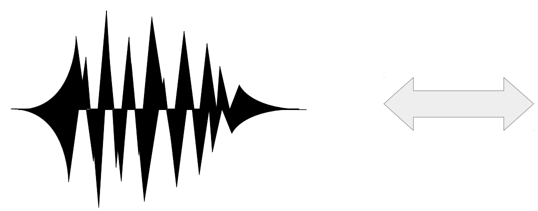
- What is style/attribute transfer
- Methods for unsupervised style transfer
- **Code:** Style transfer model

Applications 2: Dialog

he ate an apple → good, he needs to slim down

- Models for dialogs
- Ensuring diversity in outputs
- Coherence in generation
- **Code:** Play with dialog generation

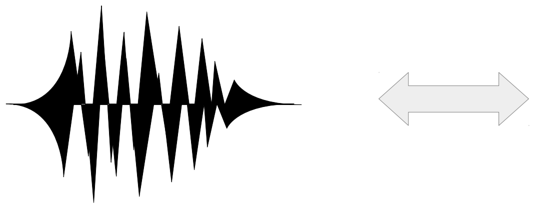
Applications 3: Speech Recognition/Synthesis



this is the song that never ends it just goes on and on my friends and if you started singing it not knowing what it was you'll just keep singing it forever just because this is the song that never ends it just goes on and on my friends and if...

- Encoding continuous sequences
- Generating continuous sequences
- **Code:** Speech recognition

Applications 4: Speech Translation

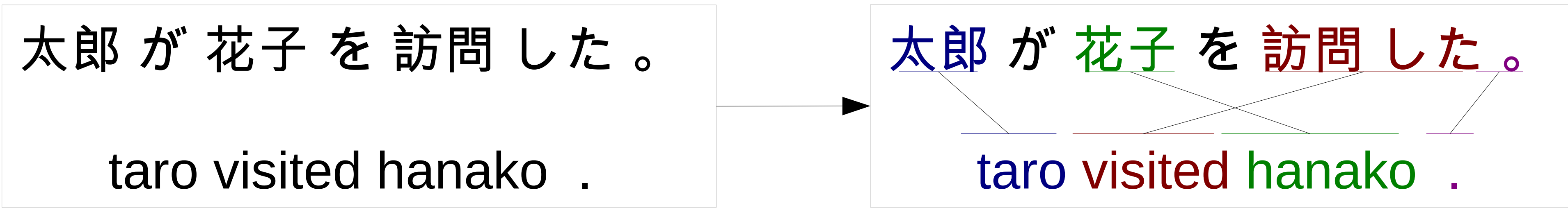


これは終わることのない歌であり、友よ、ずっと続きます。これは終わることのない歌であり、友よ、ずっと続きます。これは終わることのない歌であり、友よ、ずっと続きます。これは終わることのない歌であり、友よ、ずっと続きます。

- Considerations in speech-to-text translation
- Multi-task learning with speech recognition
- **Code:** Speech translation

Symbolic Translation Models

Symbolic Methods 1: Word Alignment



- The IBM/HMM models
- The EM algorithm
- Finding word alignments
- **Implement:** Word alignment

Symbolic Methods 2: Monotonic Transduction and FSTs

d e c i d e

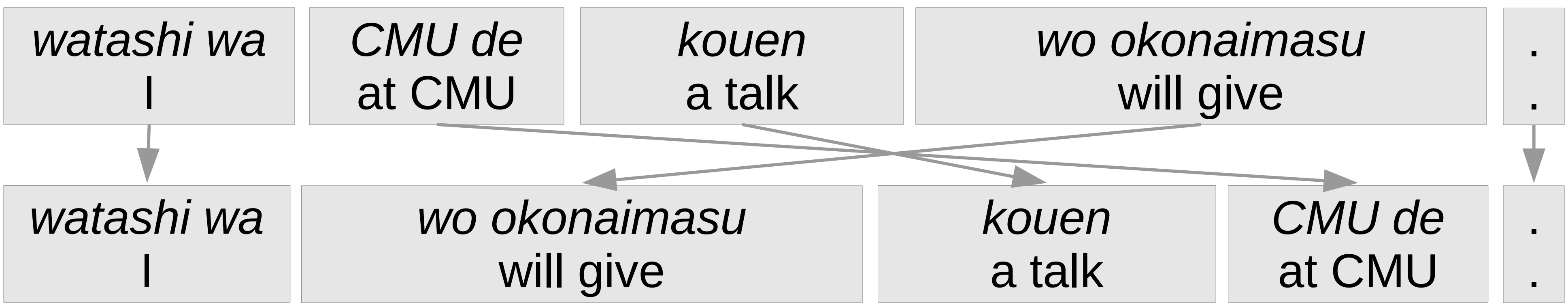
↓

u n d e c i d e d

- Models for sequence transduction
- The Viterbi algorithm
- Weighted finite-state transducers
- **Code:** A word-by-word translation model w/ FSTs

Symbolic Methods 3: Phrase-based MT

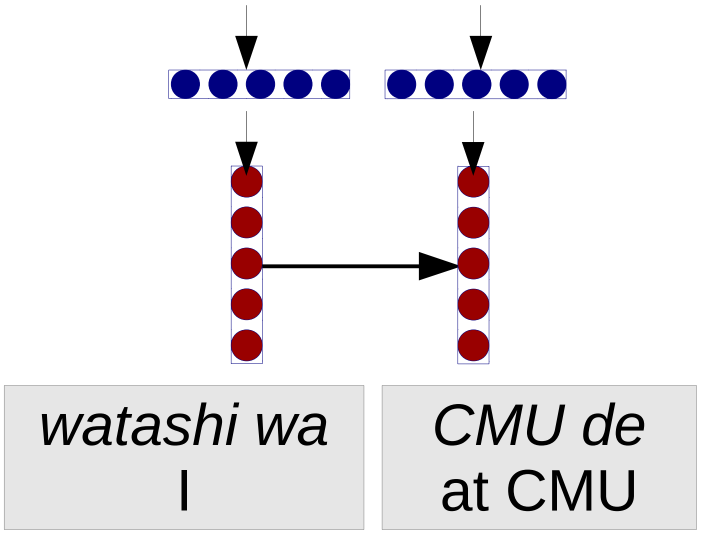
F = *watashi wa CMU de kouen wo okonaimasu .*



E = I will give a talk at CMU .

- Phrase extraction and scoring
- Reordering models
- Phrase-based decoding
- **Code:** Phrase extraction and decoding

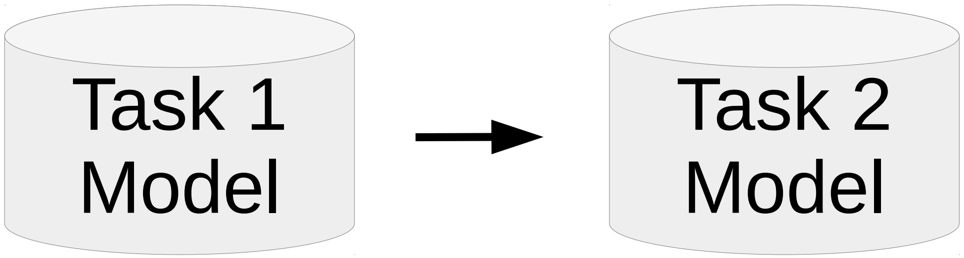
Hybrid Neural-symbolic Models



- Symbolic models with neural components
- Neural models with symbolic components
- **Code:** Implement lexicons in NMT or neural feature functions

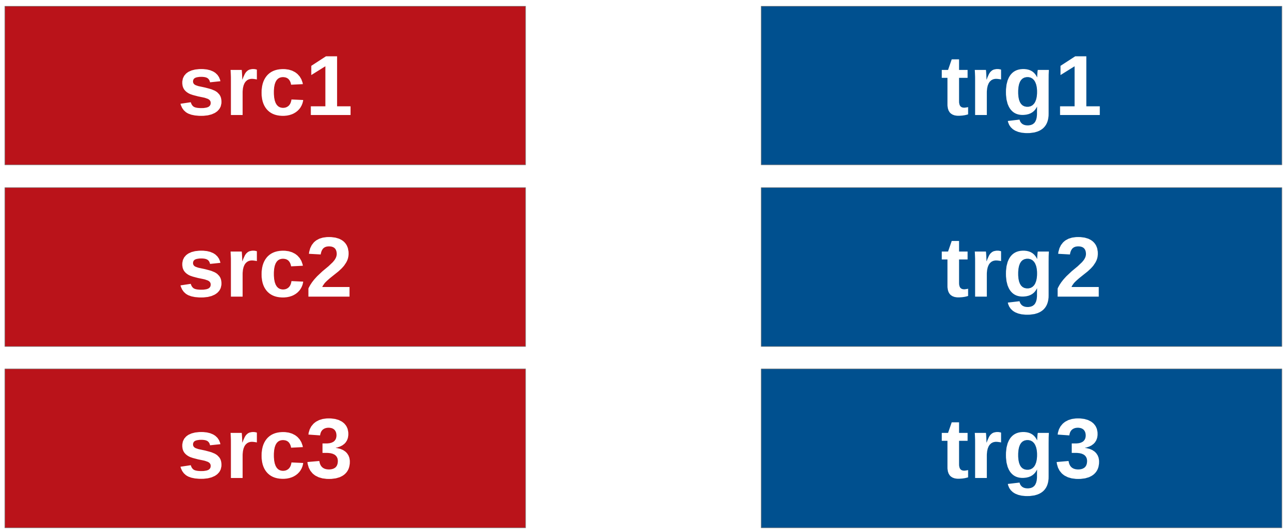
Advanced Topics

Multi-task Learning and Transfer



- Advanced multi-task learning
- Learning with other objectives
- Pre-training for seq2seq
- **Code:** Pre-training for seq2seq

Document-level Models



- Methods for capturing context
- Methods for evaluating context appropriateness
- **Code:** Context-sensitive translation

Robust Translation

Input	Luat eienr Stduie der Cambrdige Unievrstit speilt es kenie Rlloe in welcehr Reiehnfogle die Buhcstbaen in eniem Wrot vorkmomen, die eingzie whctige Sahce ist, dsas der ertse und der lettze Buhcstbaen stmimt .
Human	According to a study from Cambridge university, it doesn't matter which order letters in a word are, the only important thing is that the first and the last letter appear in their correct place.
`char2char`	Cambridge Universttte is one of the most important features of the Cambridge Universttten , which is one of the most important features of the Cambridge Universttten .
`Nematus`	Luat eienr Stduie der Cambrant Unievrstilt splashed it kenie Rlloe in welcehr Reiehnfogle the Buhcstbaen in eniem Wred vorkmomen, die eingzie whcene Sahce ist, DSAs der ertse und der lettze Buhcstbaen stmimt .

Image: Belinkov and Bisk (2017)

- Handling incorrect spelling, grammar, etc.
- Adversarial attacks
- **Code:** Adversarial attacks

For Next Class

Homework

- Read n-gram language modeling materials
- Get software working on your machine to follow along the code walks
  - By Thursday 1/19: Python and NumPy
  - By Tuesday 1/24: PyTorch and DyNet neural net libraries (use of PyTorch or DyNet is not mandatory, but examples will be in one or the other)