Machine Translation and Sequence-to-sequence Models

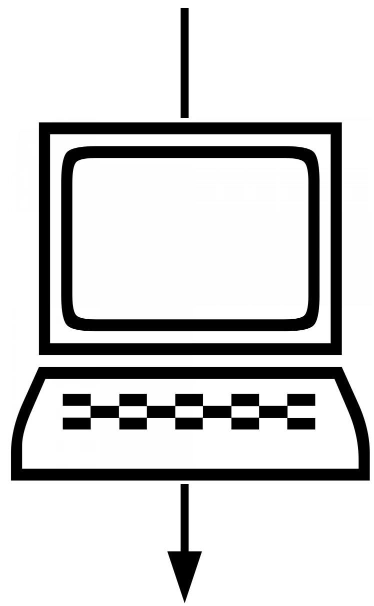
http://phontron.com/class/mtandseq2seq2018/

Graham Neubig

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

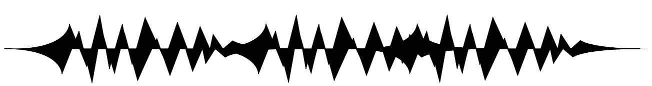
Tagging:

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

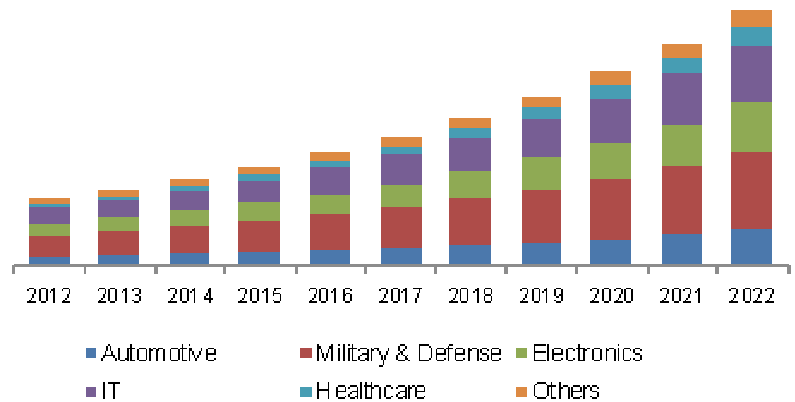
Do no evil? We literally save lives now

13 Feb 2015 at 12:01, John Leyden

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

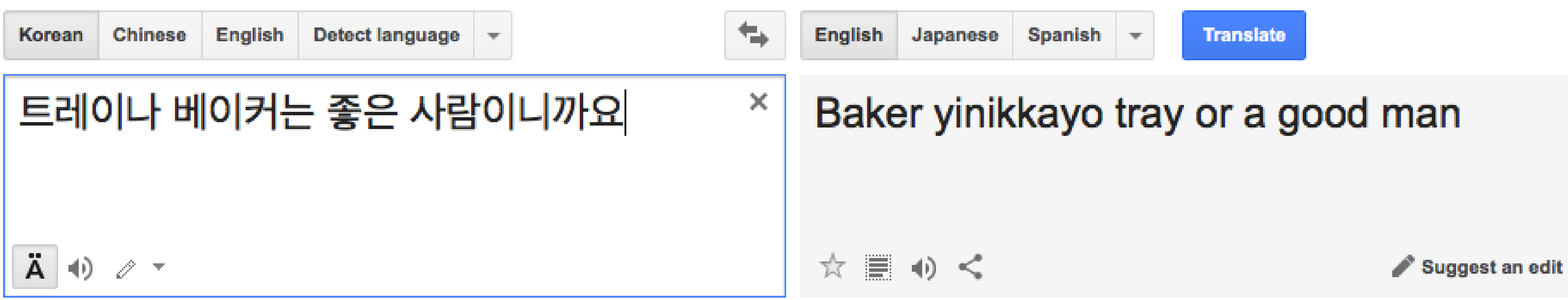
Source: The Register

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
  - Pseudo-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.
- **Final project:** Come up with an interesting new idea and test it.

Assignment Instructions

- 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.
- We recommend implementing in Python, using DyNet or PyTorch as your neural network library.

Class Grading

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

Class Plan

1. Introduction (Today): 1 class
2. Language Models: 3 classes
3. Neural MT: 3 classes
3. Evaluation/Analysis: 2 classes
4. Applications: 2 classes
5. Symbolic MT: 3 classes
7. Advanced Topics: 11 classes
8. Final Project Presentations: 2 classes

Guest Lectures

- Bob Frederking (9/13):
  Rule/Knowledge-based Translation
- Bhiksha Raj (11/27):
  Speech Applications

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?

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

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

Part 1: Neural Models

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

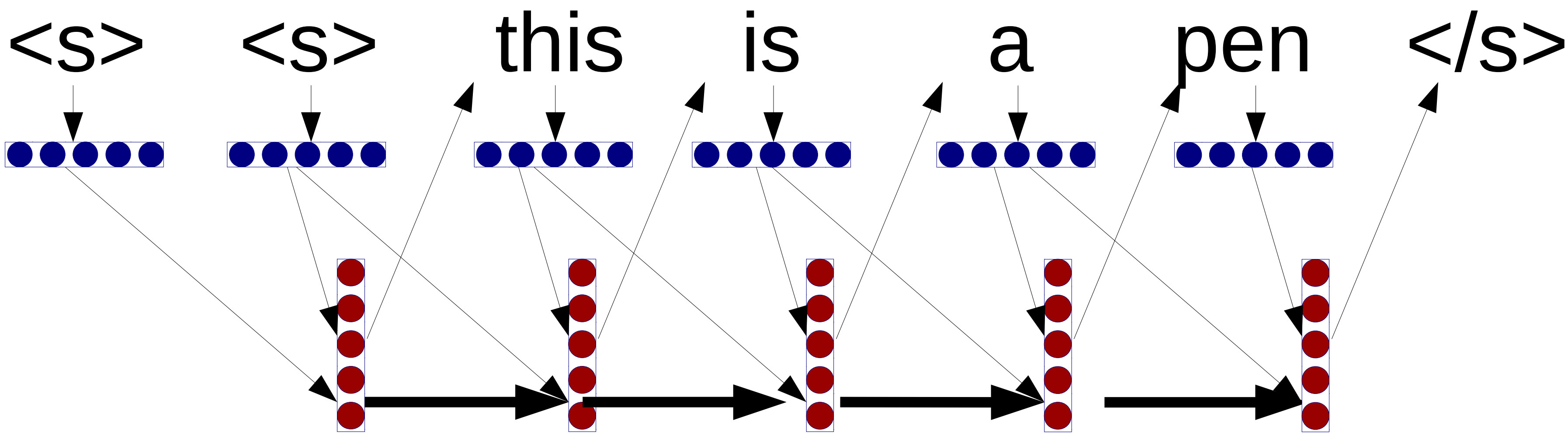
- Definition of language modeling
- Count-based n-gram language models
- Evaluating language models
- **Code Example:** n-gram language model

Language Models 2: Log-linear/Feed-forward Language Models

$$\begin{array}{l} \text{a} \\ \text{the} \\ \text{talk} \\ \text{gift} \\ \text{hat} \\ \ldots \end{array} \quad \boldsymbol{b} = \begin{pmatrix} 3.0 \\ 2.5 \\ -0.2 \\ 0.1 \\ 1.2 \\ \ldots \end{pmatrix} \quad \boldsymbol{w}_{1,a} = \begin{pmatrix} -6.0 \\ -5.1 \\ 0.2 \\ 0.1 \\ 0.6 \\ \ldots \end{pmatrix} \quad \boldsymbol{w}_{2,giving} = \begin{pmatrix} -0.2 \\ -0.3 \\ 1.0 \\ 2.0 \\ -1.2 \\ \ldots \end{pmatrix} \quad \boldsymbol{s} = \begin{pmatrix} -3.2 \\ -2.9 \\ 1.0 \\ 2.2 \\ 0.6 \\ \ldots \end{pmatrix}$$

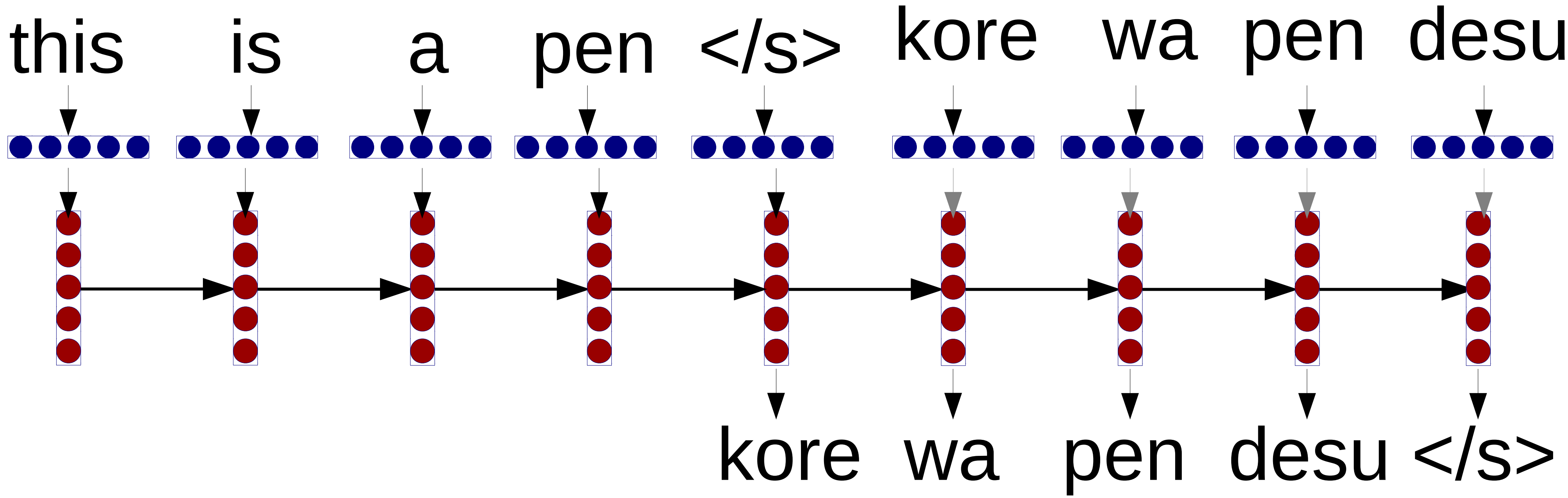
- Log-linear/feed-forward language models
- Stochastic gradient descent and mini-batching
- Features for language modeling
- **Implement:** Feed forward language model

Language Models 3: Recurrent LMs



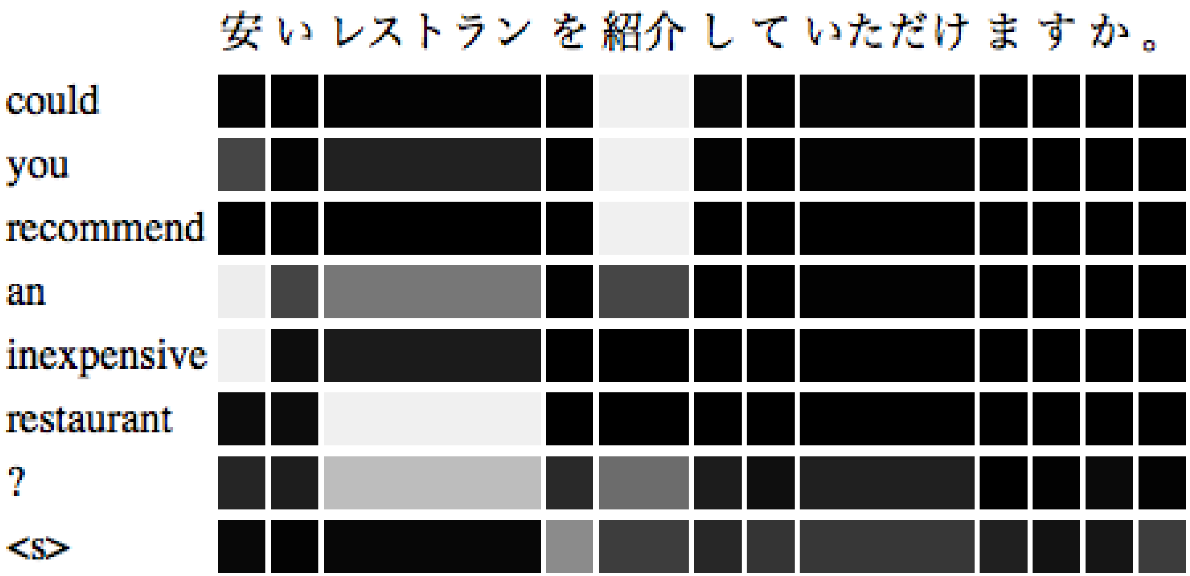
- Recurrent neural networks
- Vanishing Gradient and LSTMs/GRUs
- Regularization and dropout
- **Implement:** Recurrent neural network LM

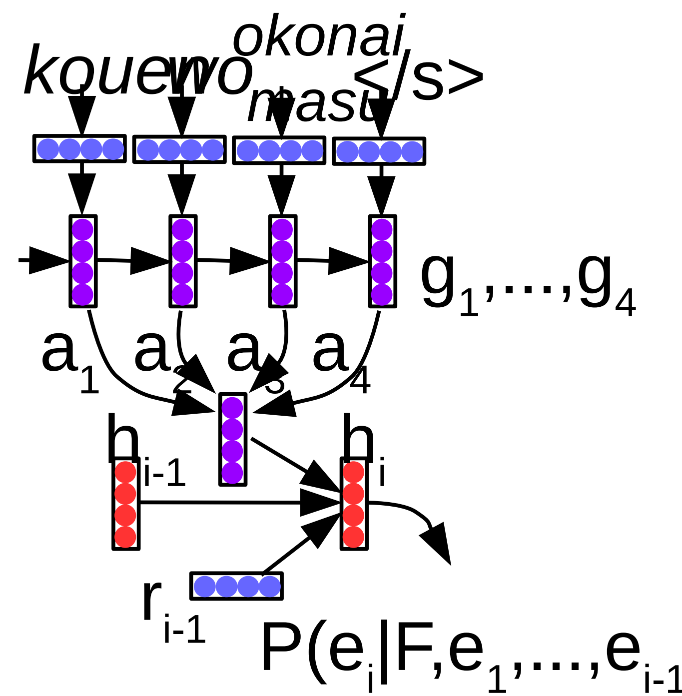
Neural MT 1: Encoder-decoder Models



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

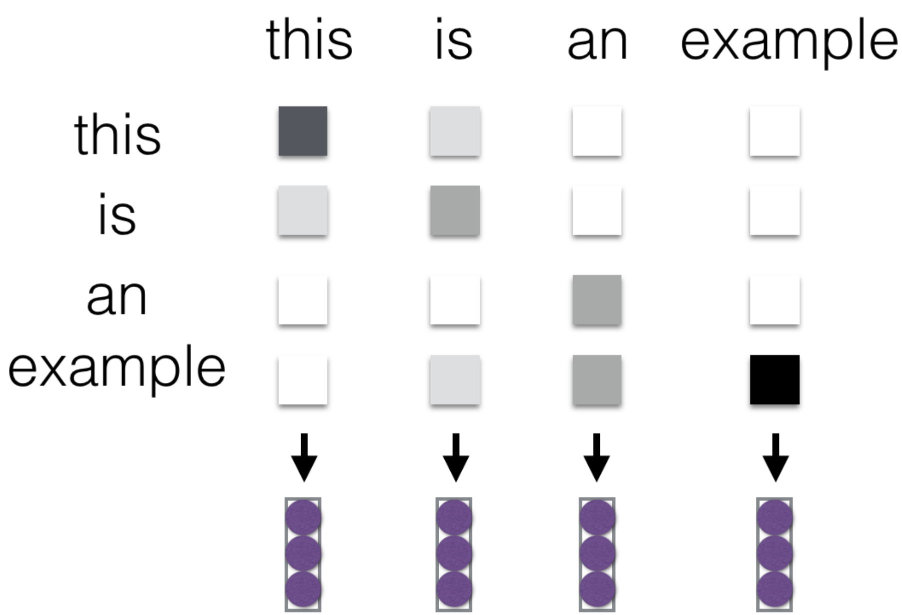
Neural MT 2: Attentional Models

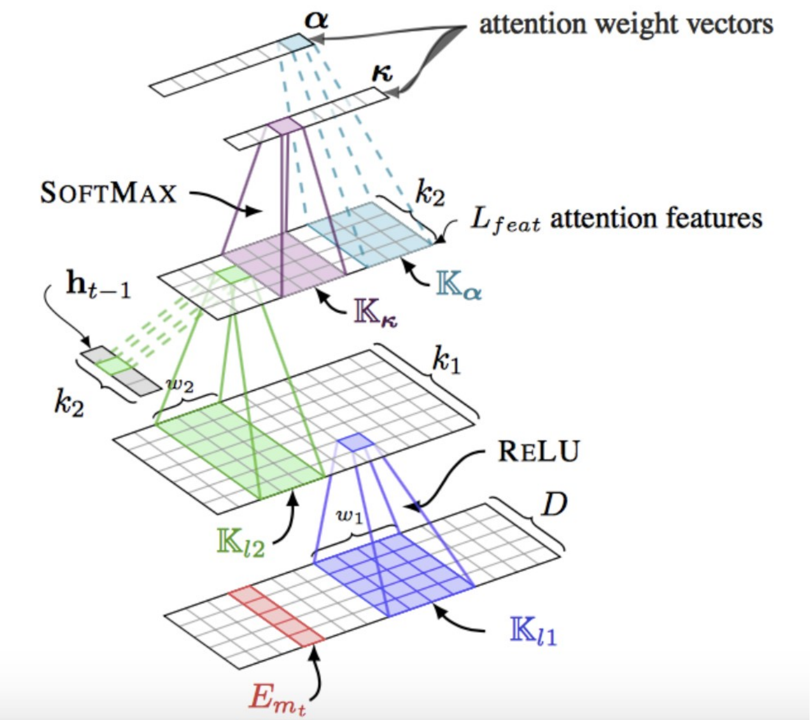




- Attention in its various varieties
- Unknown word replacement
- Attention improvements, coverage models
- **Implement:** Attentional model

Neural MT 3: Self-attention, CNNs





- Self attention
- Convolutional neural networks
- A case study, the transformer
- **Implement:** Self-attentional models

Data and Evaluation

Data/Evaluation 1a: Creating Data


毎日jp
ホーム ニュース オピニオン スポーツ エンタメ 地域 特集･連載 ENG
オピニオン 社説 余録 解説 コラム
トップ > オピニオン > 記事
[PR] 休肝日が気になる40代男性が始めた健康法！しじみ習慣／無料サンプル
ツイート 23 おすすめ 15 チェック 記事を印刷 文字サ

社説:超高齢社会　「肩車型」の常識を疑え

毎日新聞　2012年05月05日　02時30分

長寿はおめでたいことなのに、高齢化となると悲観論をもって語られることが多い。現役世代
続けているせいでもある。現役4人が高齢者1人を背負う「騎馬戦型」から、現役1人が高齢者1
車型」になると言われたら誰しも不安になるだろう。たしかに人口比率はそのようになる。

だからこそ先進国最低レベルの国民負担率（税と保険の負担）をもう少し引き上げるべきだ
「肩車型」説は登場したはずだったが、野田佳彦首相らの言い方がまずいのだろうか、逆に社

The Mainichi
[PR] ４０歳からの「しじみ習慣」休肝日が気になるあなたに!／無料サンプル
ツイート おすすめ チェック 記事を印刷 文字サイズ 小 中 大

Editorial: Aging society does not necessarily spell doom

Longevity is something to be celebrated, but when it comes to the aging of Japanese society, it is often discussed in a pessimistic tone.

One reason for this is the continuing decline in people of working age. Learning that our society is shifting from one in which four working people financially support one senior citizen, to another in which each working person must support one senior citizen -- a so-called "piggyback" setup -- would make anyone anxious. And indeed, that is exactly what is happening.

- Preprocessing
- Document harvesting and crowdsourcing
- Other tasks: dialog, captioning
- **Implement:** Find/preprocess data

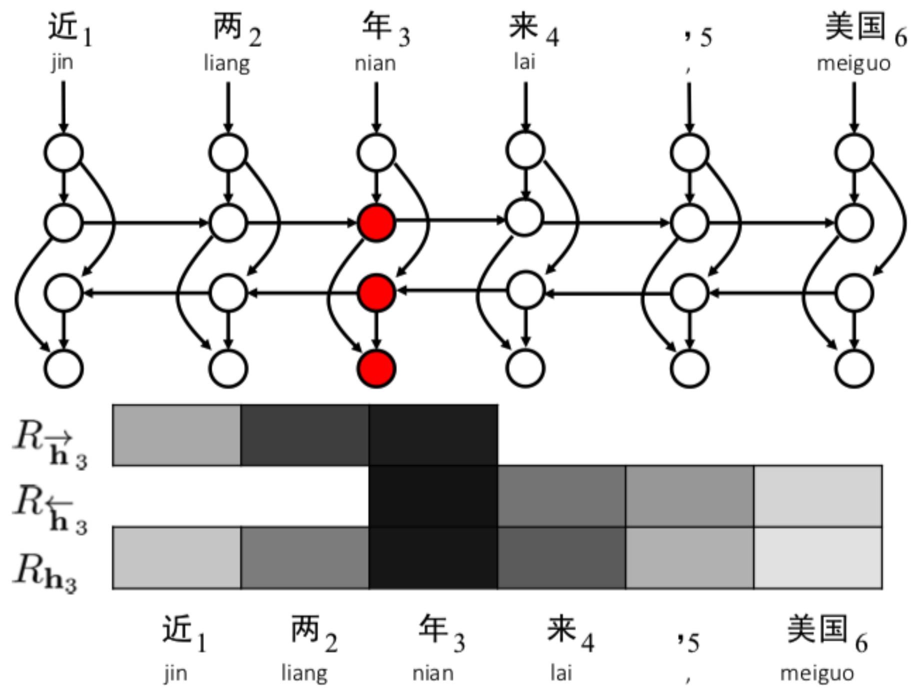
Data/Evaluation 1b: Evaluation

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
- **Implement:** BLEU and measure correlation

Data/Evaluation 2: Analysis and Interpretation



- Analyzing results
- Visualization of neural MT models
- **Implement:** Visualization of results

Application Examples

Applications 1: Summarization and Data-to-text Generation

President Trump said Monday that the United States and Mexico had reached agreement to revise key portions of the North American Free Trade Agreement and would finalize it within days, suggesting he was ready to jettison Canada from the trilateral trade pact if the country did not get on board quickly.

→ Trump Says Nafta Deal Reached Between U.S. and Mexico

- Generating shorter summaries of long texts
- Generating written summaries of data
- Necessary improvements to models
- **Implement:** Summarization model

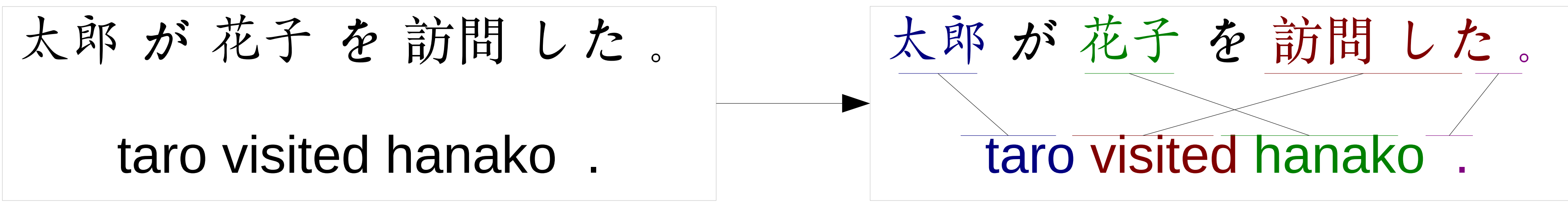
Applications 2: Dialog

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

- Models for dialogs
- Ensuring diversity in outputs
- Coherence in generation
- **Implement:** Dialog generation

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

he ate an apple

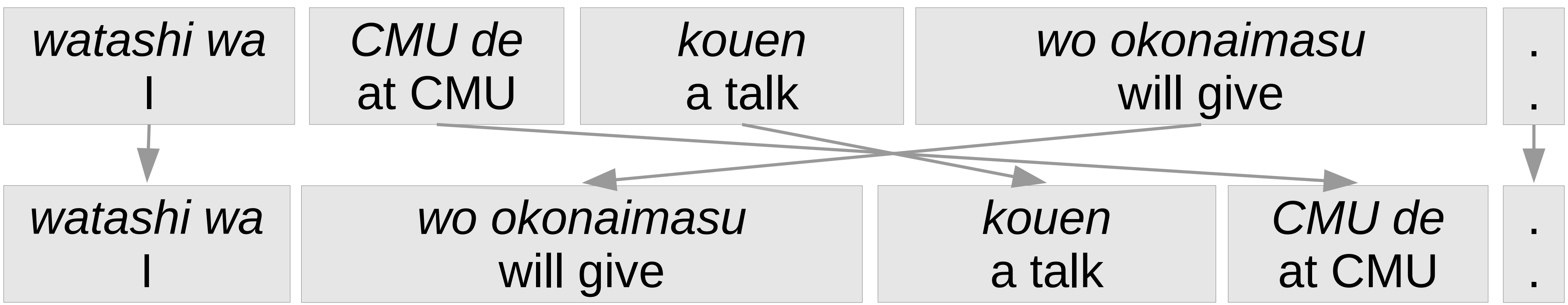
↓

PRN VBD DET PP

- Models for sequence transduction
- The Viterbi algorithm
- Weighted finite-state transducers
- **Implement:** A part-of-speech tagger

Symbolic Methods 3: Phrase-based MT

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

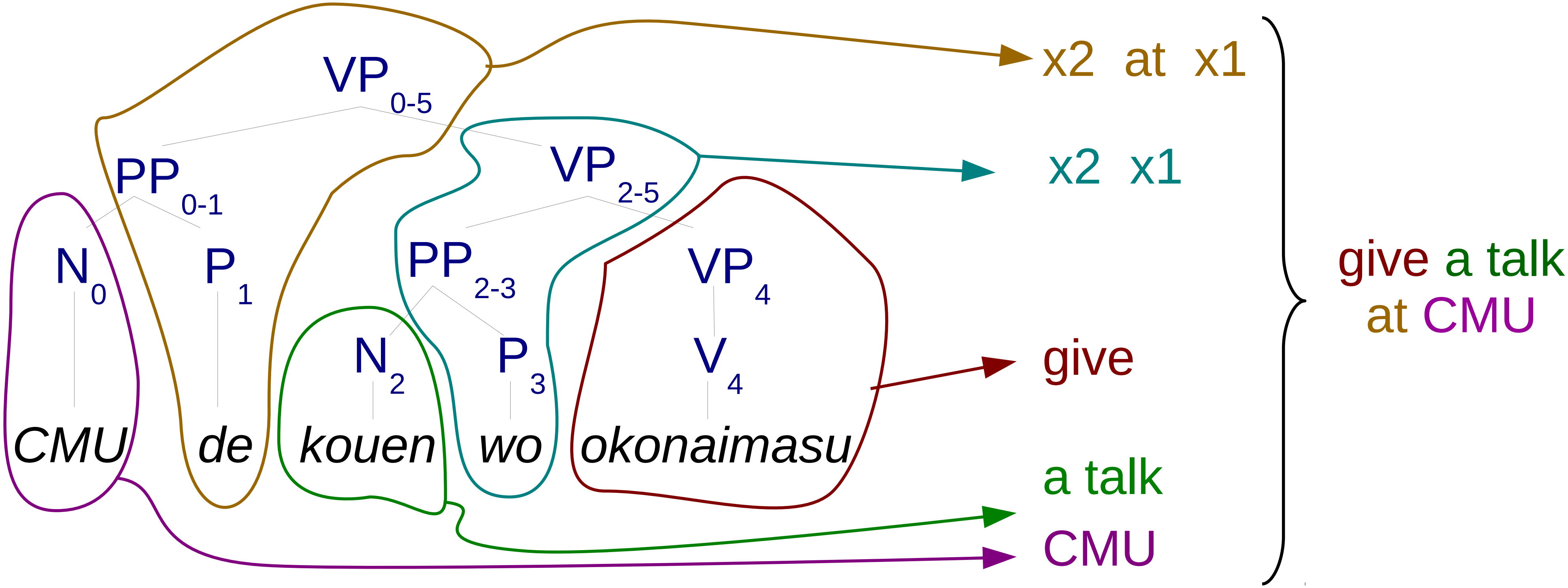


E = I will give a talk at CMU .

- Phrase extraction and scoring
- Reordering models
- Phrase-based decoding
- **Implement:** Phrase extraction or

Advanced Topics

Tree-based MT



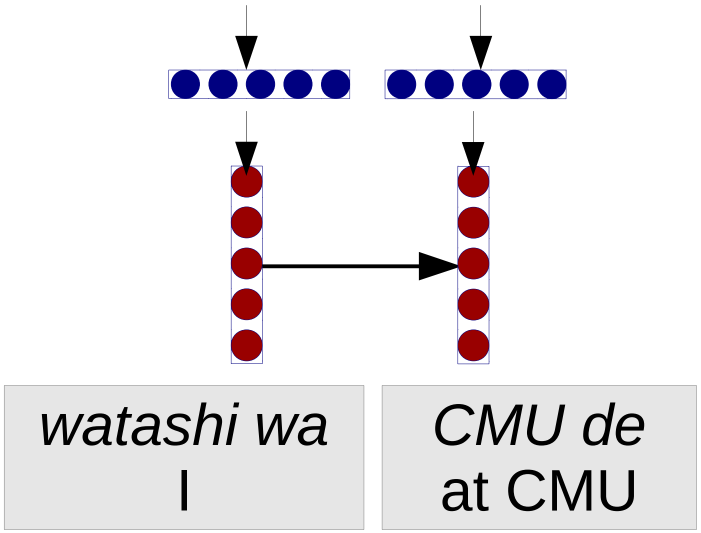
- Graphs and hyper-graphs
- Synchronous grammars
- Tree structure in neural models
- **Implement:** Tree-structured encoder

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
- **Implement:** Optimization of NMT or PBMT

Incorporating External Knowledge into NMT



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

Subword Models

reconstructed

↓

re+ construct+ ed

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

Multi-lingual and Multi-task Learning

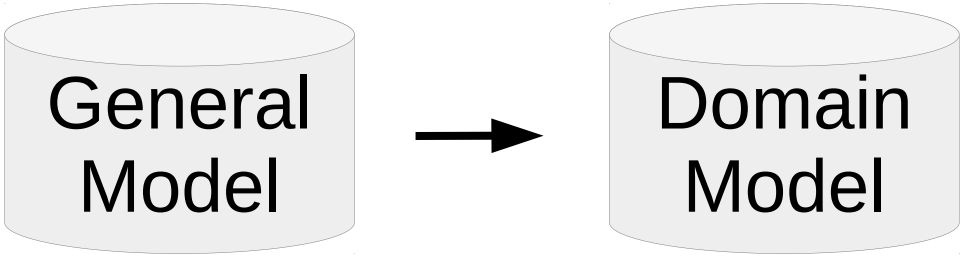
hello

こんにちは

hola

- Learning for multiple tasks
- Learning for multiple languages
- **Implement:** Implement a multi-lingual neural system

Adaptation/Transfer Learning



- Domain adaptation
- Cross-task adaptation
- **Implement:** Adaptation methods

Ensembling/System Combination

Model 1 + Model 2

- Ensembles and distillation
- Post-hoc hypothesis combination
- Reranking
- **Implement:** Ensembled decoding

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
  - By Tuesday 1/24: DyNet neural net library (use of DyNet is not mandatory for assignments, but examples will be in DyNet)