

Statistical Machine Translation

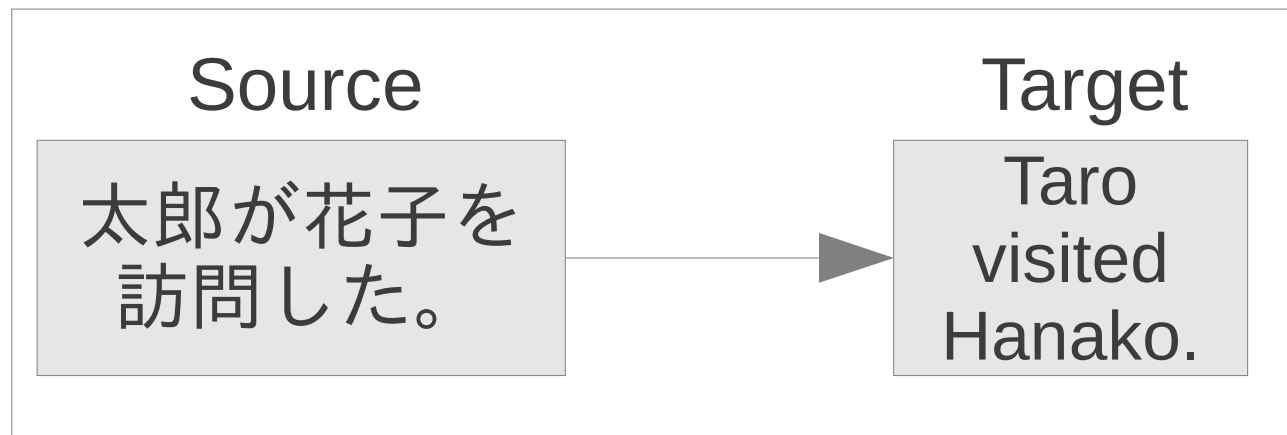
Graham Neubig

Nara Institute of Science and Technology (NAIST)

10/23/2012

Machine Translation

- Automatically translate between languages



- Real products/services being created!





NAIST Travel
Conversation
Translation System
(@AHC Lab)

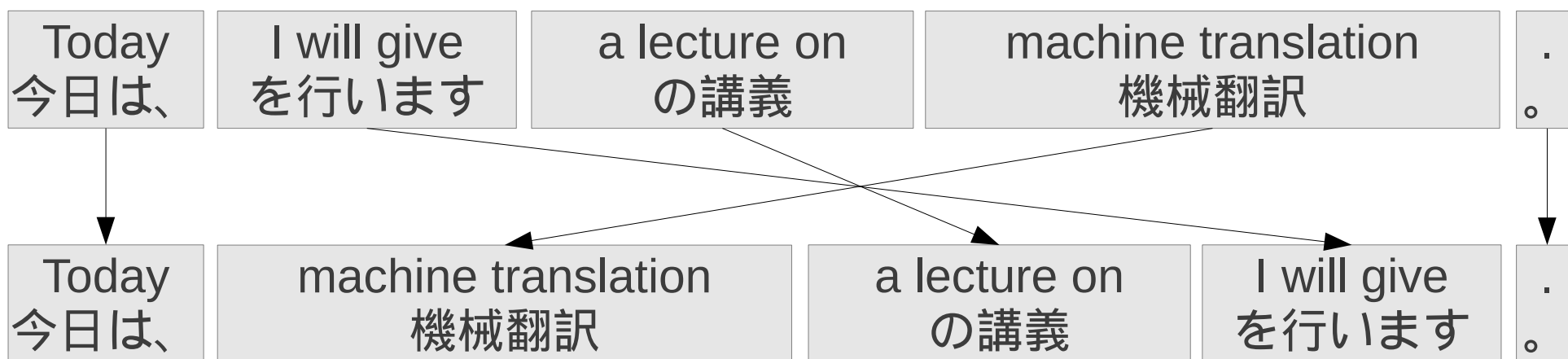
How does machine translation work?

Today I will give a lecture on machine translation .

How does machine translation work?

- Divide sentence into translatable patterns, reorder, combine

Today I will give a lecture on machine translation .



今日は、機械翻訳の講義を行います。

Problem

- There are millions of possible translations!

花子 が 太郎 に 会った

Hanako met Taro

Hanako met to Taro

Hanako ran in to Taro

Taro met Hanako

The Hanako met the Taro

- How do we tell which is better?

Statistical Machine Translation

- Translation model:

$$P(\text{“今日”} \mid \text{“today”}) = \text{high}$$

$$P(\text{“今日は、”} \mid \text{“today”}) = \text{medium}$$

$$P(\text{“昨日”} \mid \text{“today”}) = \text{low}$$

- Reordering Model:

$$P\left(\begin{array}{cc} \text{鶏} & \text{食べる} \\ \text{が} & \\ \hline \text{chicken} & \text{eats} \end{array}\right) = \text{high}$$

$$P\left(\begin{array}{cc} \text{鶏} & \text{食べる} \\ \text{を} & \\ \hline \text{eats} & \text{chicken} \end{array}\right) = \text{high}$$

$$P\left(\begin{array}{cc} \text{鶏} & \text{食べる} \\ \text{が} & \\ \hline \text{eats} & \text{chicken} \end{array}\right) = \text{low}$$

- Language Model:

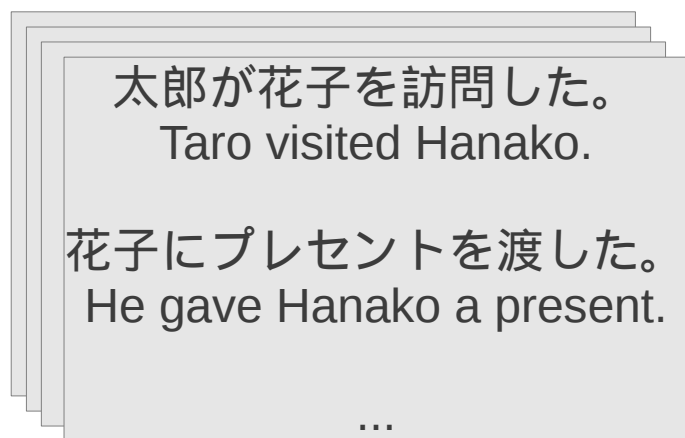
$$P(\text{“Taro met Hanako”}) = \text{high}$$

$$P(\text{“the Taro met the Hanako”}) = \text{high}$$

Creating a Machine Translation System

- Learn patterns from documents

Documents



Models

Translation Model

Reordering Model

Language Model

United Nations Text
(English/French/Chinese/Arabic ...)

Yomiuri Shimbun, Wikipedia Text
(Japanese/English)

How Do we Learn Patterns?

- For example, we go to an Italian restaurant w/ Japanese menu

チーズムース
Mousse di formaggi

タリアテッレ 4種のチーズソース
Tagliatelle al 4 formaggi

本日の鮮魚
Pesce del giorno

鮮魚のソテー お米とグリーンピース添え
Filetto di pesce su "Risi e Bisi"

ドルチェとチーズ
Dolce e Formaggi

- Try to find the patterns!

How Do we Learn Patterns?

- For example, we go to an Italian restaurant w/ Japanese menu

チーズムース

Mousse di formaggi

タリアテッレ 4種のチーズソース

Tagliatelle al 4 formaggi

本日の鮮魚

Pesce del giorno

鮮魚のソテー お米とグリーンピース添え

Filetto di pesce su “Risi e Bisi”

ドルチェとチーズ

Dolce e Formaggi

- Try to find the patterns!

Steps in Training a Phrase-based SMT System

- Collecting Data
- Tokenization
- Language Modeling
- Alignment
- Phrase Extraction/Scoring
- Reordering Models
- Decoding
- Evaluation
- Tuning

Collecting Data

- Sentence **parallel data**
 - Used in: **Translation model/Reordering model**

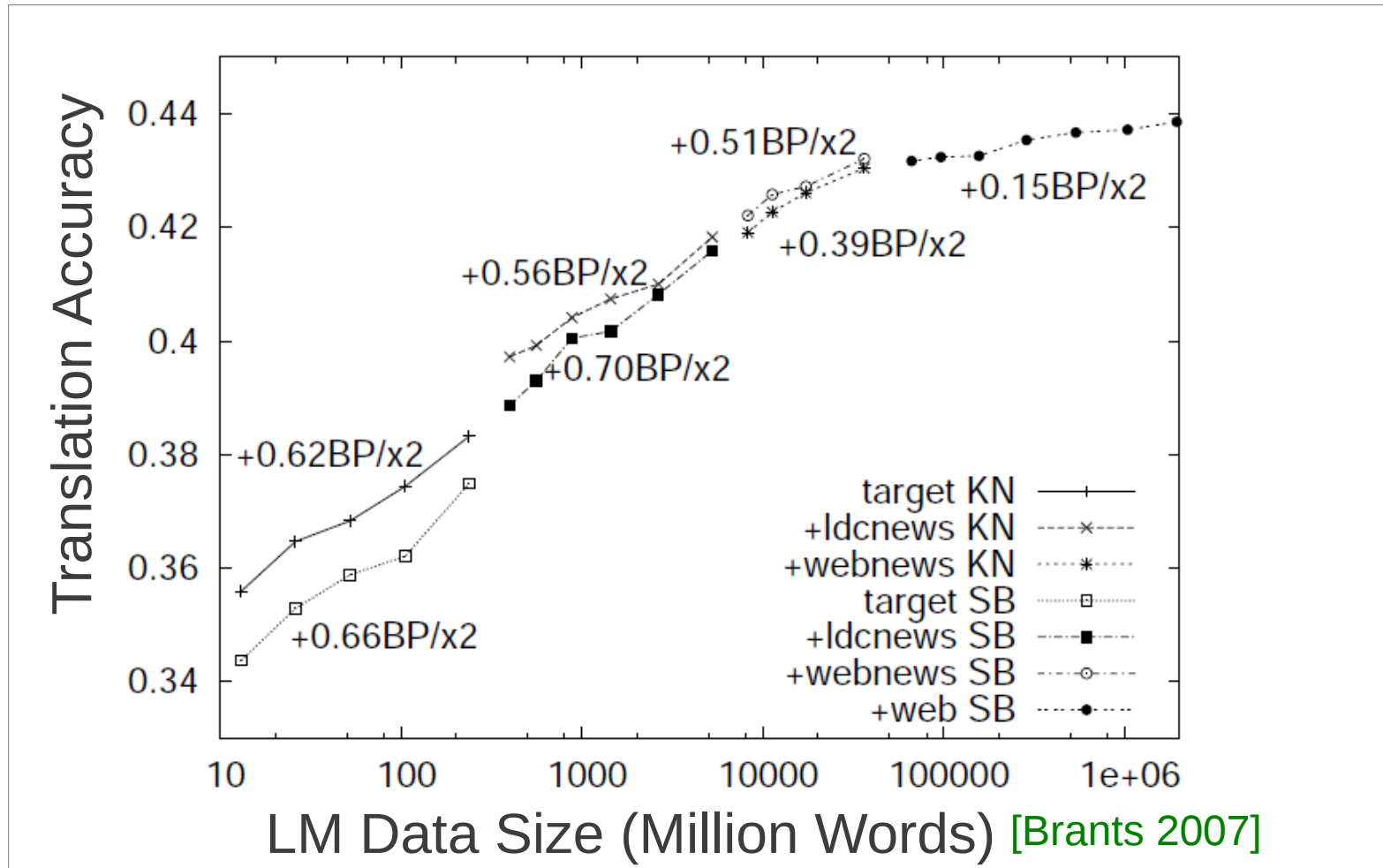
これはペンです。	This is a pen.
昨日は友達と食べた。	I ate with my friend yesterday.
象は花が長い。	Elephants' trunks are long.

- **Monolingual data** (in the target language)
 - Used in: **Language model**

This is a pen.
I ate with my friend yesterday.
Elephants' trunks are long.

Good Data is

- Big! →



- Clean
- In the same domain as test data

Collecting Data

- High quality parallel data from:
 - Government organizations
 - Newspapers
 - Patents
- Crawl the web
- Merge several data sources

Finding Data on the Web

- Find bilingual pages [Resnik 03]

毎日jp

ホーム ニュース オビニオン スポーツ エンタメ 地域 特集・連載 ENG

オビニオン 社説 余録 解説 コラム

トップ > オビニオン > 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

 0
  ツイート <23
  おすすめ <15
  チェック
  記事を印刷
 文字サ

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

 0
  ツイート <0
  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.

This unfolding state of affairs has prompted calls to raise health tax and increase water which

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。現役世代が続い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人「肩車型」になると言われたら誰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、逆に社説

Finding Data on the Web

- Finding bilingual pages [Resnik 03]
- Sentence alignment [Moore 02]

毎日jp

ホーム ニュース オビニオン スポーツ エンタメ 地域 特集・連載 ENG

オビニオン 社説 余録 解説 コラム

トップ > オビニオン > 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

 0
  23
  15
 
 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。現役世代が続い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人「肩車型」になると言われたら誰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、逆に社説

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

 0
  ツイート 0
  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.

This unfolding state of affairs has prompted calls to raise health tax and increase other contrib

Question 1:

- Write down three candidates for sources of parallel data in English-Japanese, or some other language pair you are familiar with.
- They should all be of different genres.

Tokenization

- **Example:** Divide Japanese into words

太郎が花子を訪問した。
↓
太郎 が 花子 を 訪問 した 。

- **Example:** Make English lowercase, split punctuation

Taro visited Hanako.
↓
taro visited hanako .

Tokenization is Important!

- **Just Right:** Can translate properly

太郎	が	→	taro	○
太郎	を	→	taro	○

- **Too Long:** Cannot translate if not in training data

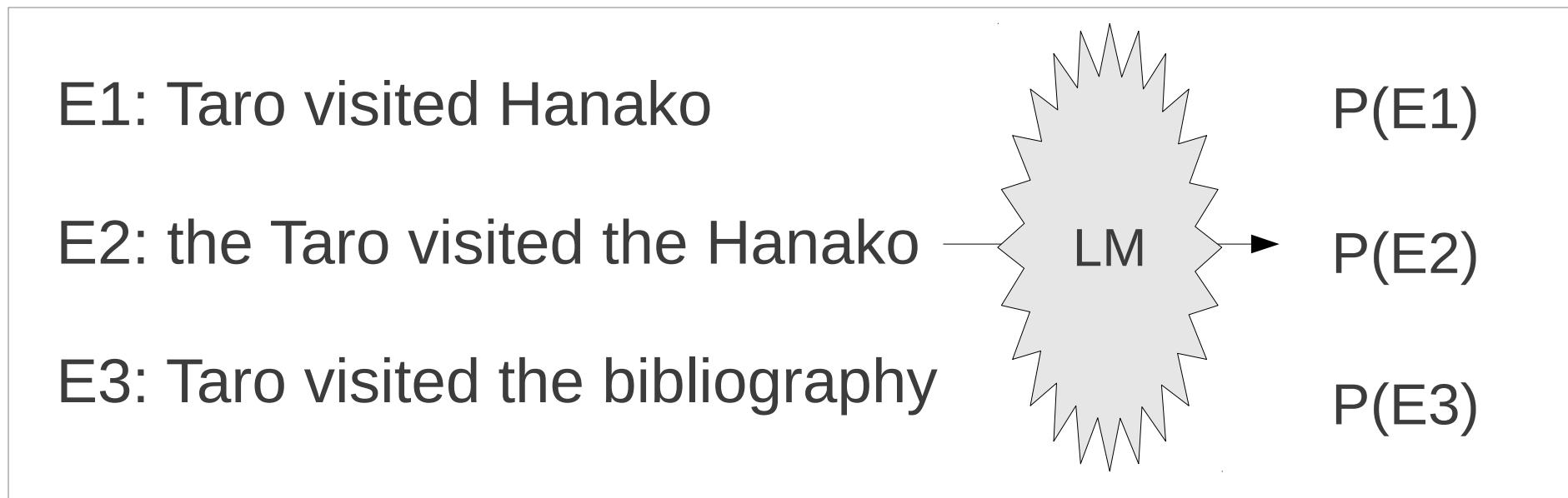
太郎が	→	taro	○	In Data
太郎を	→	太郎を	×	Not in Data

- **Too Short:** May mistranslate

太郎	が	→	fat ro	×
太郎	を	→	fat ro	×

Language Modeling

- Assign a probability to each sentence



- More fluent sentences get higher probability

$$P(E1) > P(E2)$$

$$P(E1) > P(E3)$$

n-gram Models

- We want the probability of

$$P(W = \text{"Taro visited Hanako"})$$

- **n-gram model** calculates one word at a time
 - Condition on n-1 previous words
e.g. 2-gram model

$$\begin{aligned} &P(w_1 = \text{"Taro"}) * P(w_2 = \text{"visited"} \mid w_1 = \text{"Taro"}) \\ &\quad * P(w_3 = \text{"Hanako"} \mid w_2 = \text{"visited"}) \\ &\quad * P(w_4 = \text{"</s>"} \mid w_3 = \text{"Hanako"}) \end{aligned}$$

NOTE:

sentence ending symbol </s>

Calculating n-gram Models

- n-gram models are estimated from data:

$$P(w_i | w_{i-n+1} \dots w_{i-1}) = \frac{c(w_{i-n+1} \dots w_i)}{c(w_{i-n+1} \dots w_{i-1})}$$

i live in **osaka** . </s>

i am a graduate student . </s>

my school is in **nara** . </s>

$$n=2 \rightarrow P(\text{osaka} | \text{in}) = c(\text{in osaka})/c(\text{in}) = 1 / 2 = 0.5$$

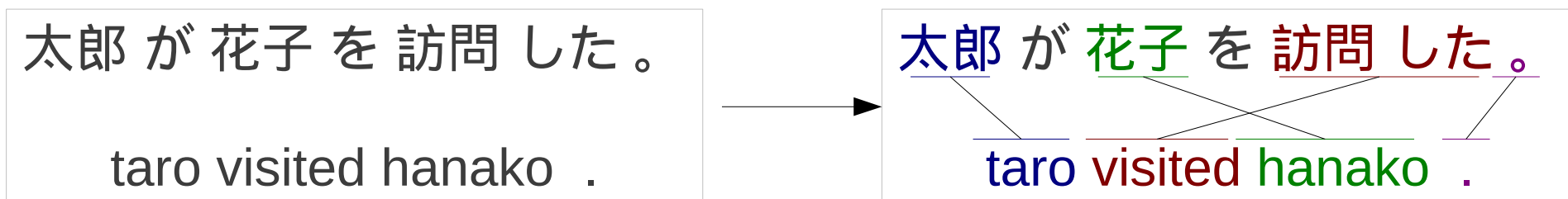
$$P(\text{nara} | \text{in}) = c(\text{in nara})/c(\text{in}) = 1 / 2 = 0.5$$

Question 2:

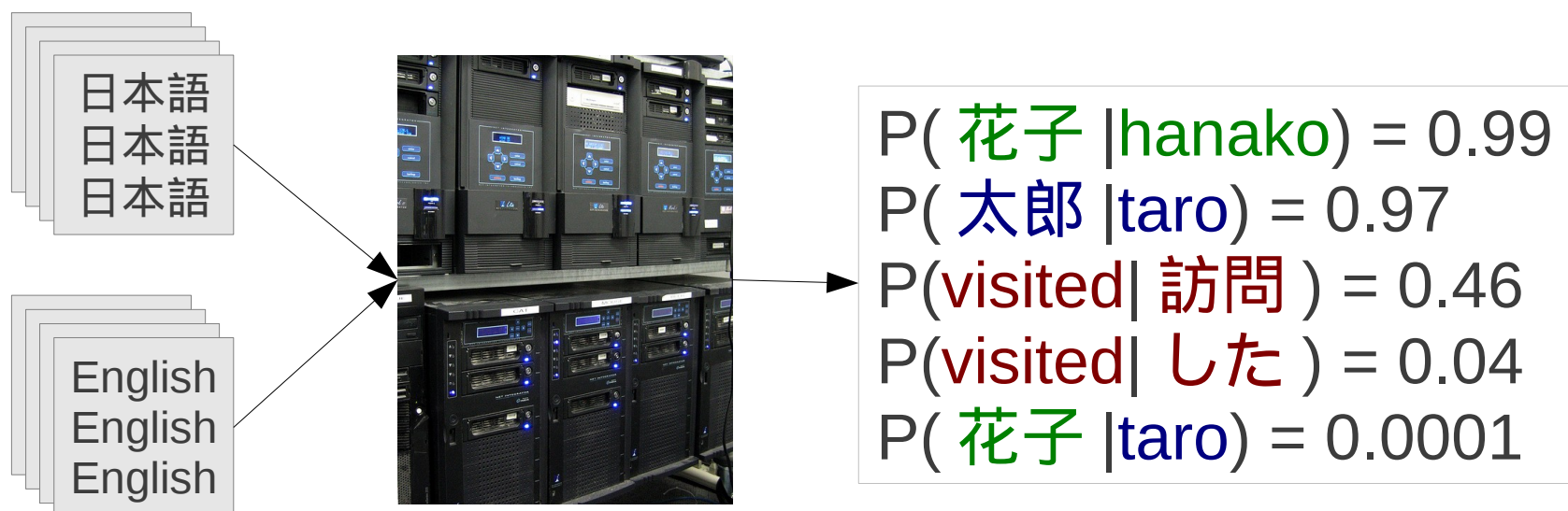
- Calculate the 2-gram probabilities of the n-grams on the worksheet.

Alignment

- Find which words correspond to each-other

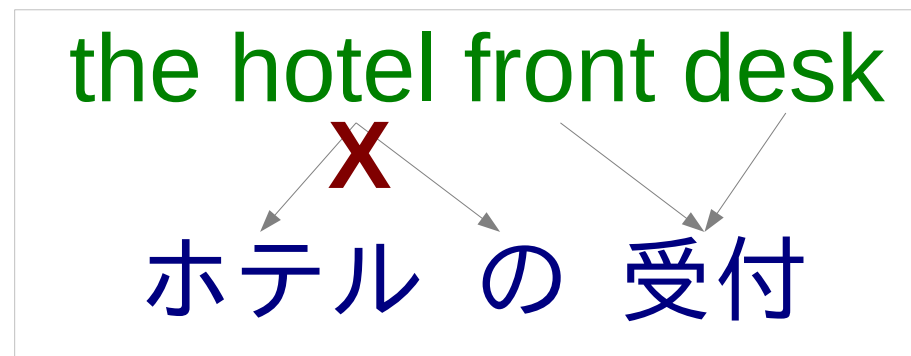
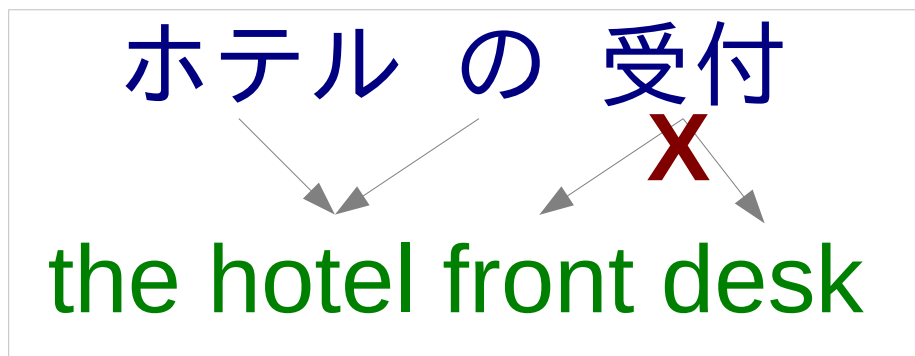


- Done automatically with probabilistic methods



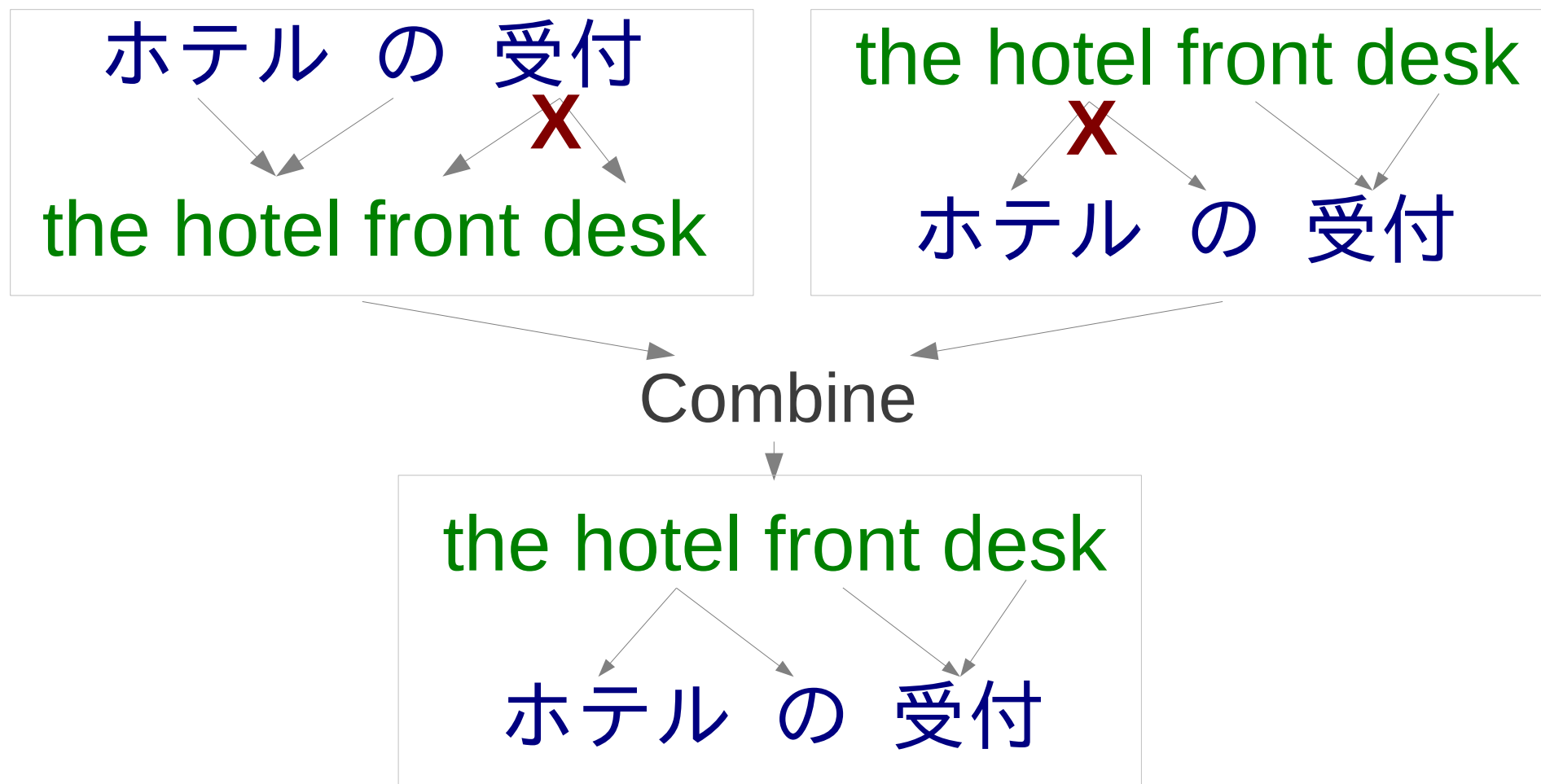
IBM/HMM Models

- One-to-many alignment model



- IBM Model 1: No structure (“bag of words”)
- IBM Models 2-5, HMM: Add more structure

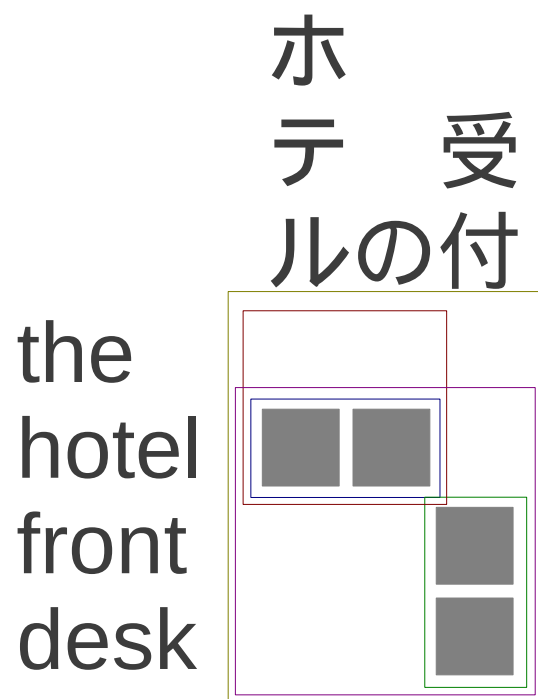
Combining One-to-Many Alignments



- Several different heuristics

Phrase Extraction

- Use alignments to find phrase pairs



ホテルの → hotel

ホテルの → the hotel

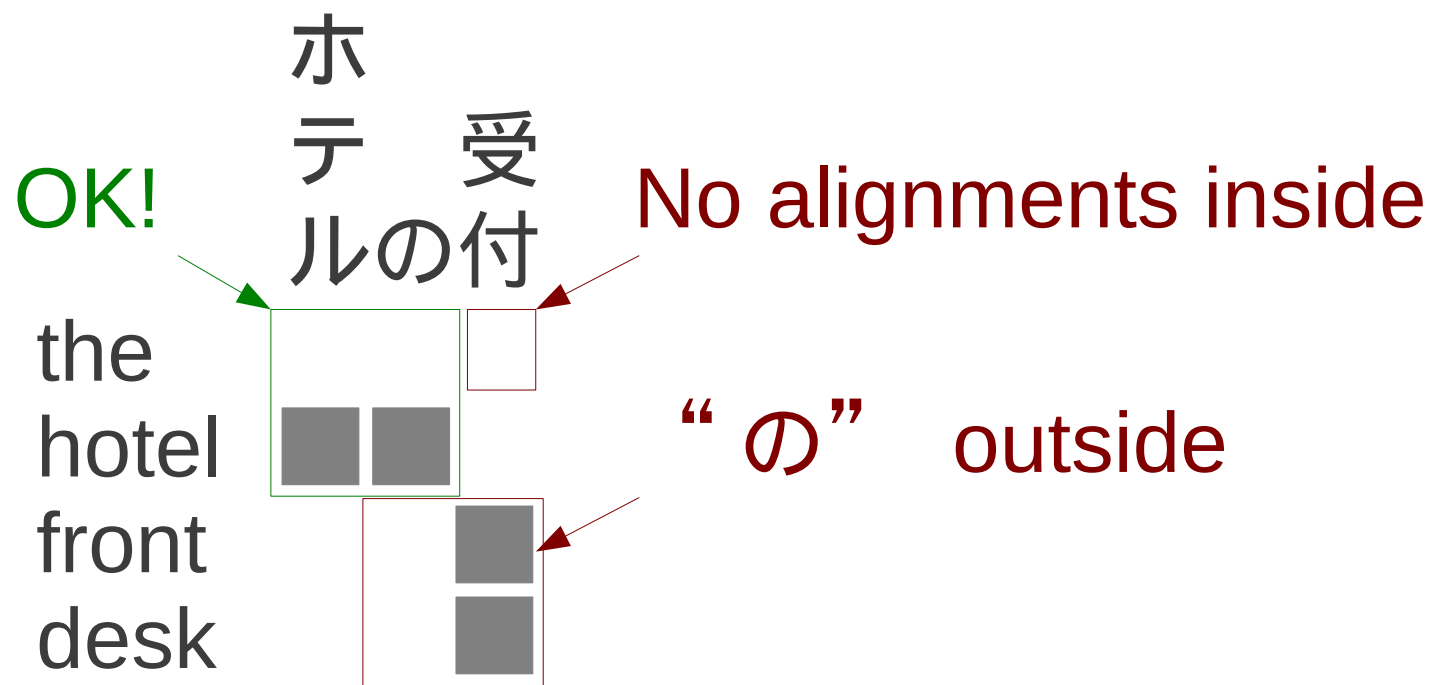
受付 → front desk

ホテルの受付 → hotel front desk

ホテルの受付 → the hotel front desk

Phrase Extraction Criterion

- Must have
 - 1) one alignment inside the phrase
 - 2) no alignments outside and in the same row/column



Question 3:

- Given the alignments on the work sheet, which phrases will be extracted by the machine translation system?

Phrase Scoring

- Calculate 5 standard features

- **Phrase Translation Probabilities:**

$$P(\mathbf{f}|\mathbf{e}) = c(\mathbf{f},\mathbf{e})/c(\mathbf{e}) \quad P(\mathbf{e}|\mathbf{f}) = c(\mathbf{f},\mathbf{e})/c(\mathbf{f})$$

e.g. $c(\text{ホテル の}, \text{the hotel}) / c(\text{the hotel})$

- **Lexical Translation Probabilities**

- Use word-based translation probabilities (IBM Model 1)
- Helps with sparsity

$$P(\mathbf{f}|\mathbf{e}) = \prod_f \frac{1}{|\mathbf{e}|} \sum_e P(\mathbf{f}|\mathbf{e})$$

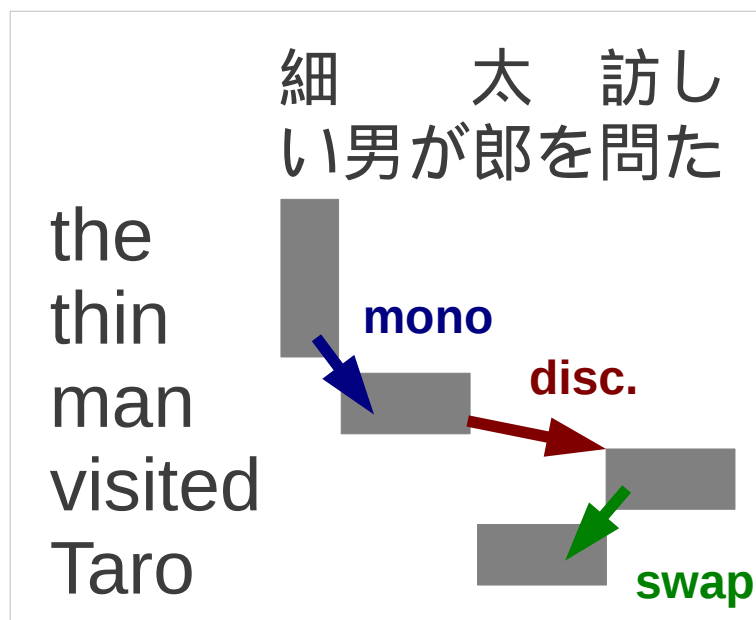
e.g.

$(P(\text{ホテル}|\text{the})+P(\text{ホテル}|\text{hotel}))/2 * (P(\text{の}|\text{the})+P(\text{の}|\text{hotel}))/2$

- **Phrase penalty:** 1 for each phrase

Lexicalized Reordering

- Probability of monotone, swap, discontinuous



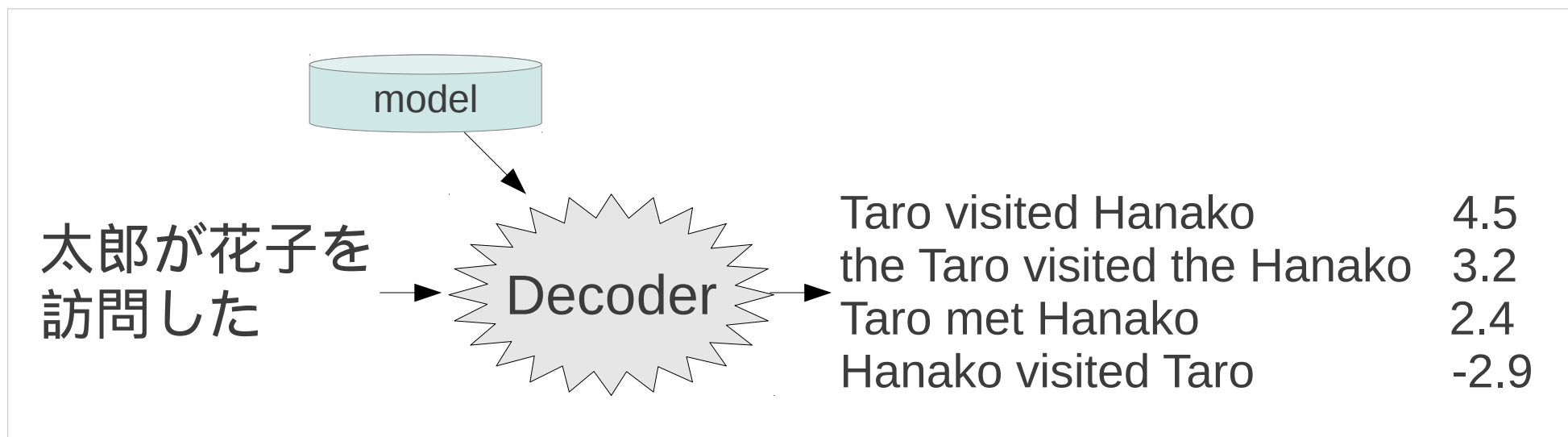
細い → the thin
high **monotone** probability

太郎 を → Taro
high **swap** probability

- Conditioning on input/output, left/right, or both

Decoding

- Given the models, **find the best answer** (or n-best)



- Exact search is NP-hard! [Knight 99]
- Decoding uses beam-search to find an approximate solution [Koehn 03]

Phrase-Based Decoding

- Build translation from left to right
- Remember which words were already translated
- Choose translation with highest score

en: he visited the white house

ja:

en: he visited the white house

ja: 彼は

en: he visited the white house

ja: 彼は ホワイトハウスを

en: he visited the white house

ja: 彼は ホワイトハウスを 訪問した

Question 4:

- How would a phrase-based machine translation system generate the translation on the work sheet?

Evaluation

- We built a machine translation system, we need to know:
 - **How good** is our system?
 - Is system A **better than** system B?
 - What are the **problems** with our system?

Human Evaluation

- **Adequacy:** Is the meaning correct?
- **Fluency:** Is the sentence natural?
- **Pairwise:** Is X a better translation than Y?

太郎が花子を訪問した


 Taro visited Hanako the Taro visited the Hanako Hanako visited Taro

Adequate?	○	○	×
Fluent?	○	×	○
Better?	B, C	C	

Automatic Evaluation

- How well does the translation match a reference?
 - (or multiple references: more than one correct translation)
- **BLEU**: n-gram precision, brevity penalty [Papineni 03]

Reference: Taro visited Hanako

System: the Taro visited the Hanako

1-gram: 3/5

2-gram: 1/4

Brevity: $\min(1, |\text{System}|/|\text{Reference}|) = \min(1, 5/3)$

brevity penalty = 1.0


$$\begin{aligned} \text{BLEU-2} &= (3/5 * 1/4)^{1/2} * 1.0 \\ &= 0.387 \end{aligned}$$

- Also **METEOR** (normalizes synonyms), **TER** (# of changes), **RIBES** (reordering)

Tuning


- **Scores** of translation, reordering, and language models

	<u>LM</u>	<u>TM</u>	<u>RM</u>	
○ Taro visited Hanako	-4	-3	-1	-8
✗ the Taro visited the Hanako	-5	-4	-1	-10
✗ Hanako visited Taro	-2	-3	-2	-7

Best Score ✗ 

- If we **add weights**, we can get better answers:

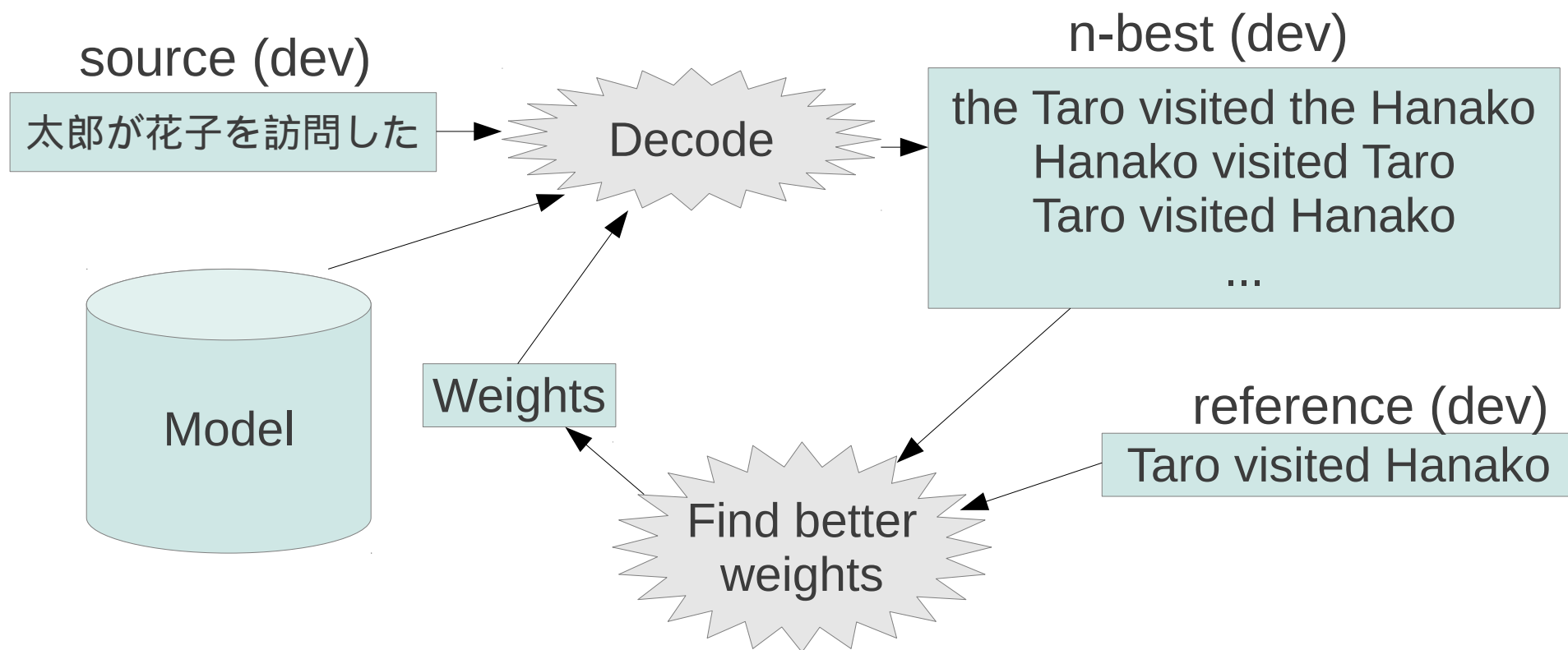
	<u>LM</u>	<u>TM</u>	<u>RM</u>	
○ Taro visited Hanako	$0.2^* -4$	$0.3^* -3$	$0.5^* -1$	-2.2
✗ the Taro visited the Hanako	$0.2^* -5$	$0.3^* -4$	$0.5^* -1$	-2.7
✗ Hanako visited Taro	$0.2^* -2$	$0.3^* -3$	$0.5^* -2$	-2.3

Best Score ○ 

- Tuning finds these weights: $w_{LM} = 0.2$ $w_{TM} = 0.3$ $w_{RM} = 0.5$

Tuning Methods

- Minimum error rate training: MERT [Och 03]



- Others: MIRA [Watanabe 07] (online update), PRO (ranking) [Hopkins 11]

Question 5:

- Given the list of hypotheses on the worksheet, find weights that maximize the BLEU score.

Assignment

Assignment (choose one):

- Paraphrasing Sentences
 - a) Use Google Translate to find at least 10 sentences that are not translated properly, and guess why.
 - b) Create a strategy to paraphrase the sentences so they are easier to translate. Explain why this strategy works.
- Manual Evaluation:
 - a) Using provided translation results, perform a manual Adequacy/Fluency evaluation, report the distribution of scores (1-5) and some examples of good/bad scores.
 - b) For 5-10 bad translations, discuss why translation failed.
- Creating a Translation System:
 - a) Follow the steps on this page to make a machine translation system and measure the accuracy:
<http://www.statmt.org/moses/?n=Moses.Baseline>
 - b) Find a setting of the system that changes the accuracy, and discuss its effect.

Assignment Submission

- Use the **additional materials** on the web site if you choose the “Manual Evaluation” assignment.
- **Length:** 500+ words plus figures/tables if necessary
- **Address:** neubig@is.naist.jp
- **Deadline:** 2012-10-30, 23:59