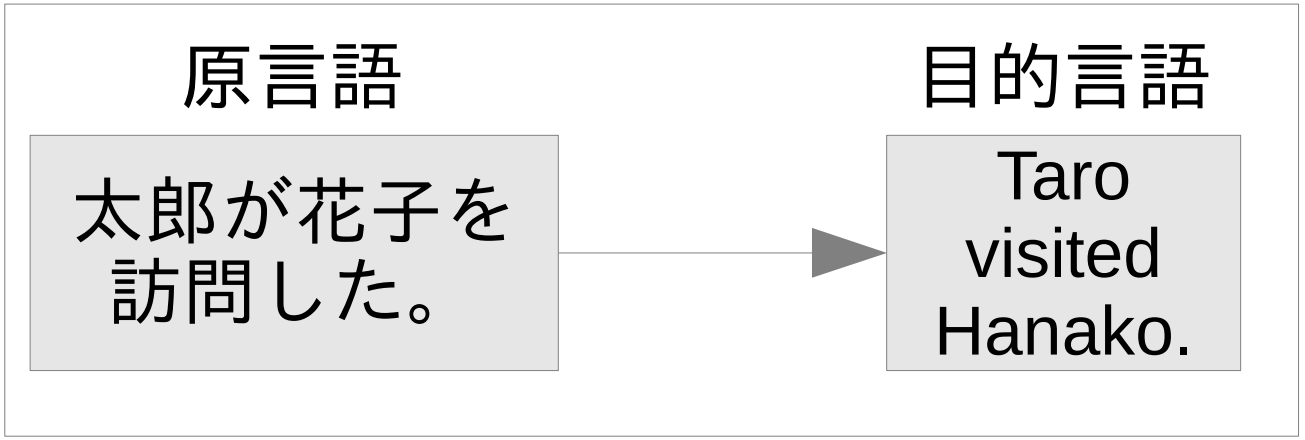
機械翻訳

～なぜできなかったのか？なぜできるようになりつつあるのか？～

Graham Neubig
奈良先端科学技術大学院大学　(NAIST)
2014-05-25

機械翻訳

- 原言語から目的言語へと自動的に翻訳



- 近年に著しい発展と実用化

Google translate

excite

Institute of Science and Technology Nara NAIST ®

KYOTO UNIVERSITY KYOTO JAPAN FOUNDED 1897

機械翻訳実験に参加した人間翻訳者の言葉 [Green+ 13]

「業務で Google や Babel を使ったことがあるけれど、あなたの翻訳システムの方がよっぽど精度が良い！」

実は Google 翻訳だった…

なぜここまでできるようになったのか？

1) モデルの構築に統計手法を用いる　統計的機械翻訳
2) 単語列ごとに翻訳を行う　フレーズベース機械翻訳
3) 翻訳の評価を効率化する　自動評価
4) 翻訳を最適化問題として扱う　識別学習
5) 文の構造を考慮する　統語ベース翻訳

1. 統計的機械翻訳

翻訳システムの人手構築

- 人手によって変換ルールを構築

辞書項目
英語の「 white 」を日本語の「白い」へ変換
英語の「 run 」を日本語の「走る」へ変換

並べ替えルール
英語の「主語・動詞・目的語」を日本語の「主語は・目的語を・動詞」

問題！

- 人手によるコスト
  （日本語・中国語？日本語・タイ語？）
- 例外

white day →
- ✗ 白い日
- ○ ホワイトデー

he has blue eyes →
- ✗ 彼は青い目を持っている
- ○ 彼の目は青い

統計的機械翻訳 [Brown+ 93]

- 大量の学習データからシステムを自動的に学習

対訳文

太郎が花子を訪問した。
Taro visited Hanako.

花子にプレゼントを渡した。
He gave Hanako a present.

...

モデル

- 翻訳モデル
- 並べ替えモデル
- 言語モデル

統計的機械翻訳

- 翻訳モデル (TM)

P(“ 今日” |“today”) = high　　P(“ 今日　は　、　” |“today”) = medium

P(“ 昨日” |“today”) = low

- 並べ替えモデル (RM)

P(鶏　が　食べる → chicken　eats)=high　P(鶏　を　食べる → eats　chicken)=high　P(鶏　が　食べる → eats　chicken)=low

- 言語モデル (LM)

P(“Taro met Hanako”)=high　P(“the Taro met the Hanako”)=low

パターン学習の例

- 日本語メニューのあるイタリア料理屋に行ったら

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

- 「formaggi」の日本語訳は？「pesce」？「e」？

パターン学習の例

- 日本語メニューのあるイタリア料理屋に行ったら

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

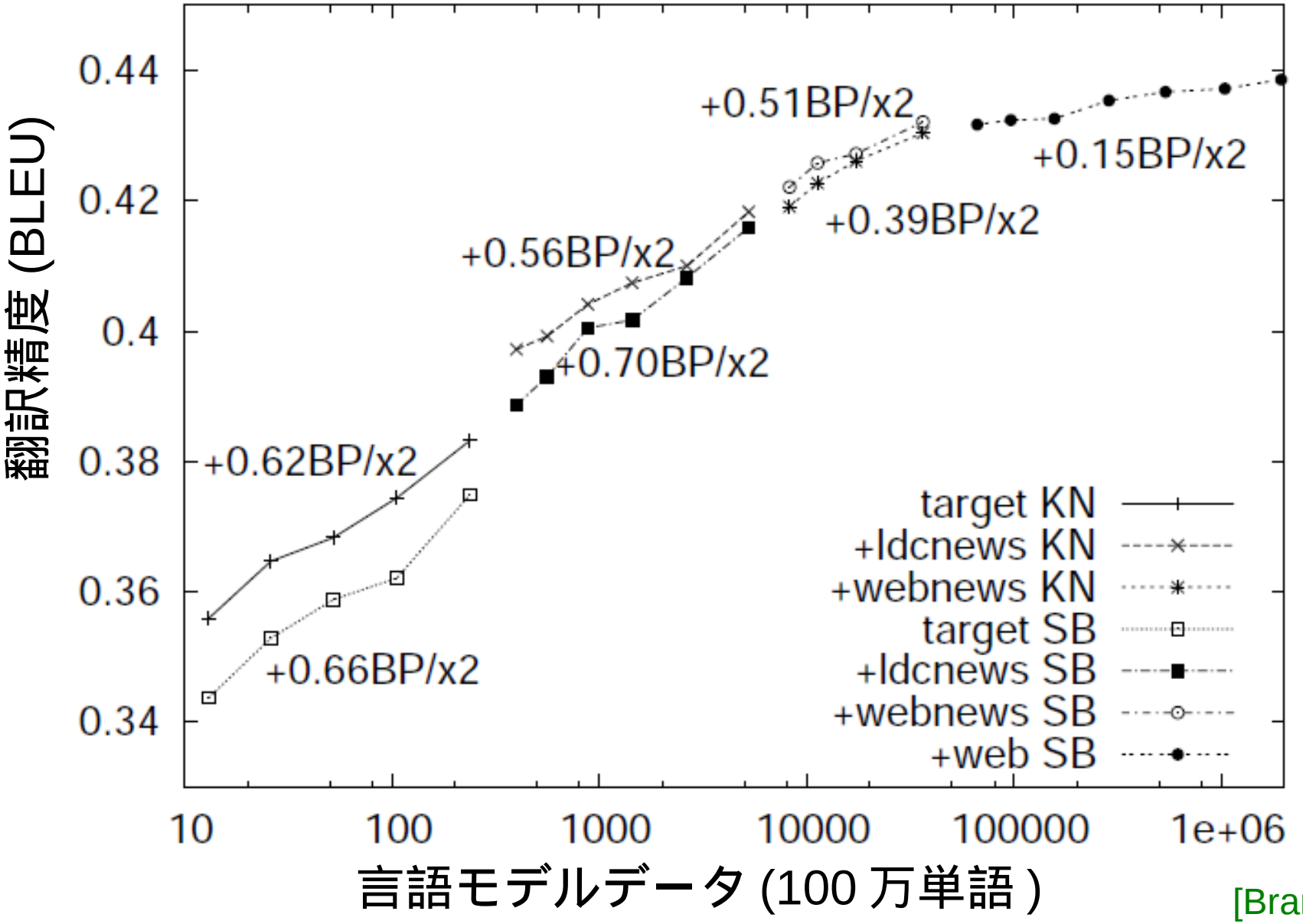
本日の鮮魚
Pesce del giorno

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

- 「formaggi」の日本語訳は？「pesce」？「e」？

データの威力



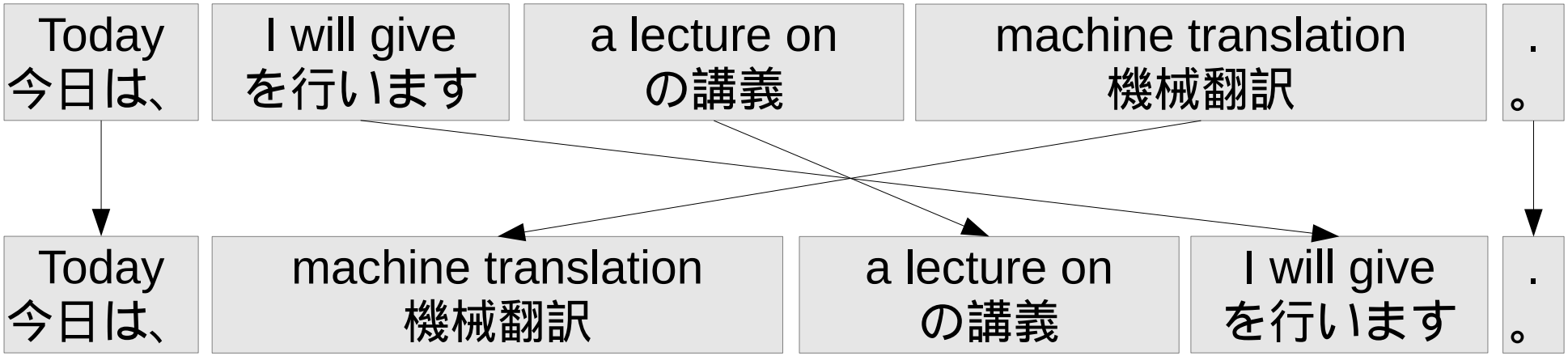
[Brants 2007]

2. フレーズベース機械翻訳

フレーズベース機械翻訳 [Koehn+ 03]

- 文をフレーズ（単語列）ごとに翻訳して、並べ替え

Today I will give a lecture on machine translation .



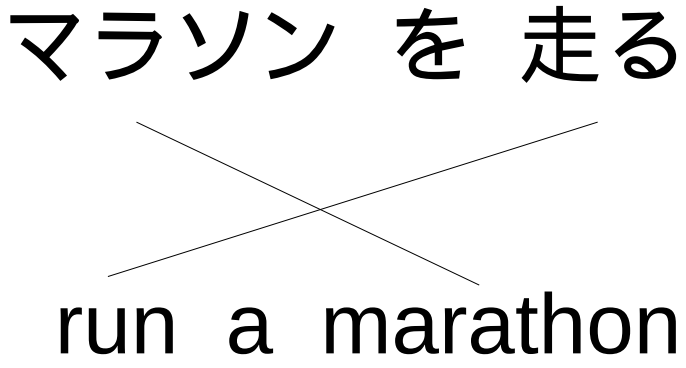
今日は、機械翻訳の講義を行います。

なぜフレーズ？

	原言語フレーズ	目的言語フレーズ	スコア
通常辞書に含まれる単語	white	白	0.5
	white	白い	0.25
	white	白色	0.1
	white	ホワイト	0.05
通常辞書に含まれない単語	white	白　を	0.05
	white	白　が	0.05
複合語	white rice	白米	1.0
	white day	ホワイト　・　デー	1.0
	white house	ホワイト　・　ハウス	1.0

フレーズの自動獲得

- 単語の対応に基づいてフレーズを列挙



marathon → マラソン

run → 走る

a marathon → マラソン
a marathon → マラソン を
marathon → マラソンを

run a → 走る
run a → を 走る
run → を 走る

run a marathon → マラソンを走る

3. 自動評価

人手評価

- 意味的妥当性 (Adequacy): 原言語文の意味が伝わるか
- 流暢性 (Fluency): 目的言語文が自然か
- 比較評価 (Pairwise): X と Y どっちの方が良いか

- 精度は良いが、非常に労力がかかる

自動評価

- システム出力は正解文に一致するか

正解文 :	Taro visited Hanako
システム出力 :	the Taro visited the Hanako

- 翻訳の正解は単一ではないため、複数の正解を利用することも

正解文 :	Taro visited Hanako Taro ran in to Hanako
システム出力 :	the Taro visited the Hanako

BLEU [Papineni+ 03]

- 最初に提案された自動評価尺度
- 未だに最も標準的
- n-gram 適合率 + 短さペナルティ

Reference: Taro visited Hanako

System: the Taro visited the Hanako

1-gram: 3/5
2-gram: 1/4
brevity penalty = 1.0

Brevity: min(1, |System|/|Reference|) = min(1, 5/3)

BLEU-2 = $(3/5*1/4)^{1/2} * 1.0$
= 0.387

4. 識別学習

翻訳の最適化 [Och+ 02]

- 各モデルのスコアを組み合わせた解のスコア

	LM	TM	RM	
○ Taro visited Hanako	-4	-3	-1	-8
X the Taro visited the Hanako	-5	-4	-1	-10
X Hanako visited Taro	-2	-3	-2	-7 ↘ 最大 X

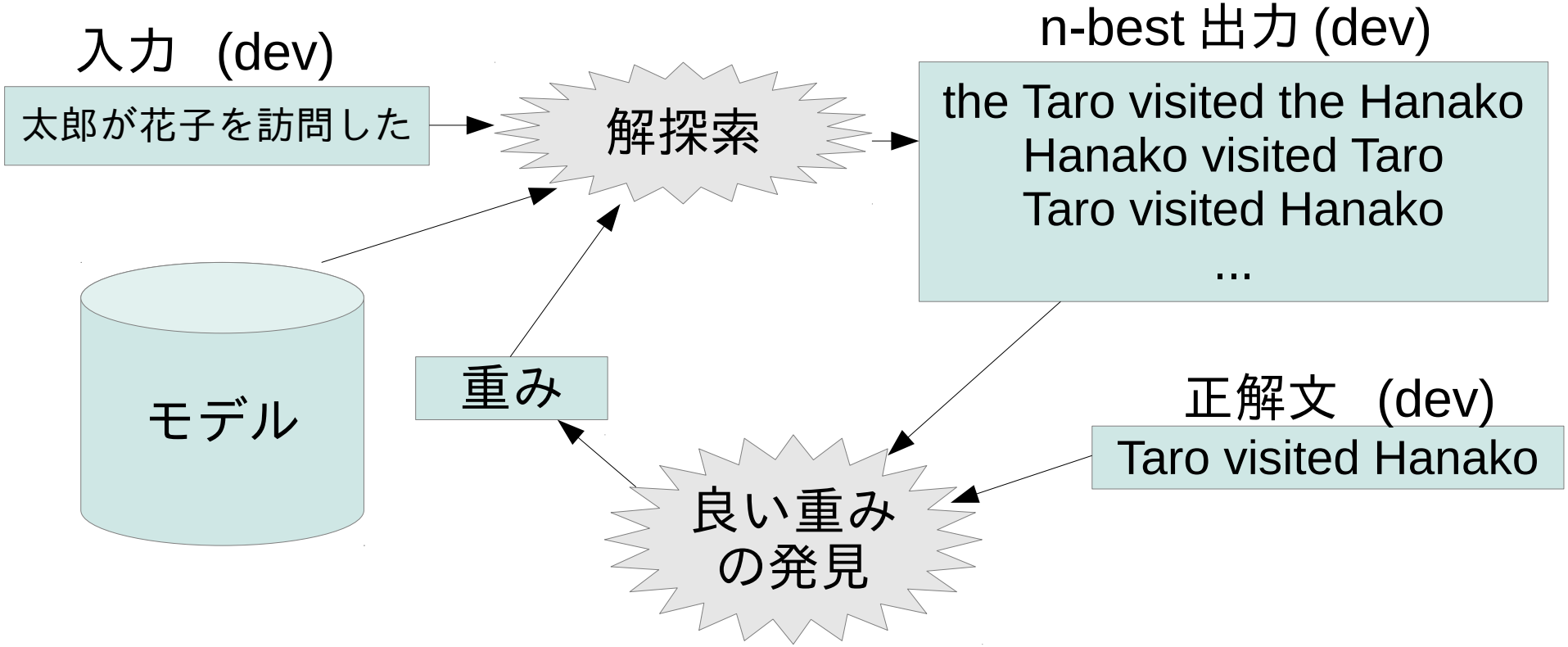
- スコアを重み付けると良い結果が得られる

	LM	TM	RM	
○ Taro visited Hanako	0.2*-4	0.3*-3	0.5*-1	-2.2 ↙ 最大 ○
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

- 最適化は重みを発見 : $w_{LM}=0.2$ $w_{TM}=0.3$ $w_{RM}=0.5$

翻訳の最適化の難しさ

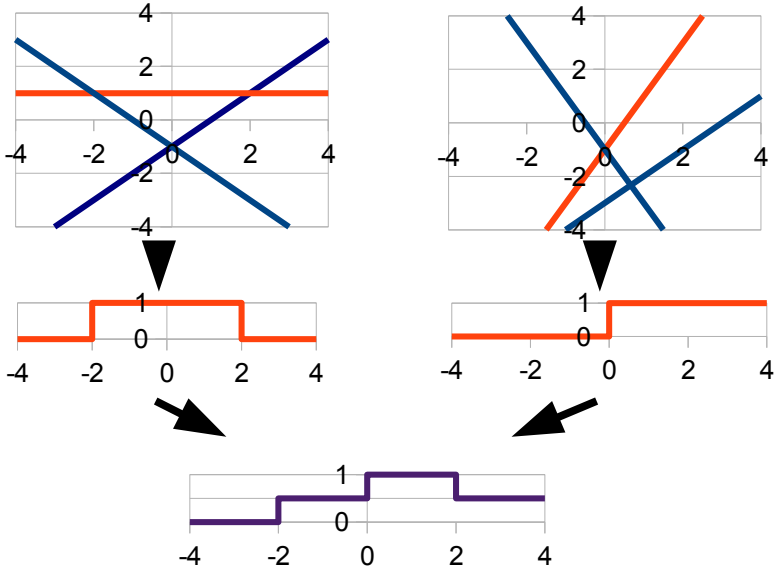
- 膨大な仮説空間 → 仮説の列挙が難しい



- BLEU などの不安定性 → 一般化が難しい

機械翻訳のための最適化手法

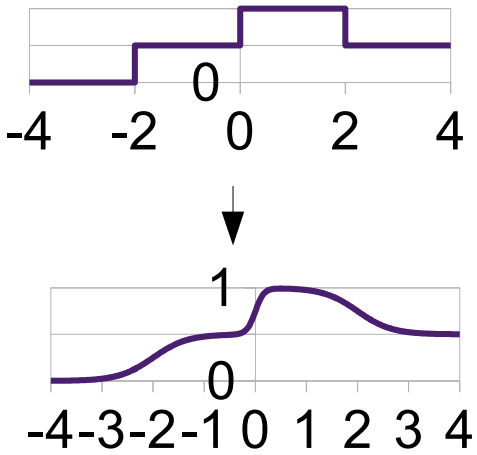
誤り率最小化学習 [Och+ 04]



オンライン学習 [Watanabe+ 07]

$$\boldsymbol{w}_t = \boldsymbol{w}_{t-1} + \varphi\left(e_i^*\right) - \varphi\left(\hat{e}_i\right)$$

勾配法 [Smith+ 06]

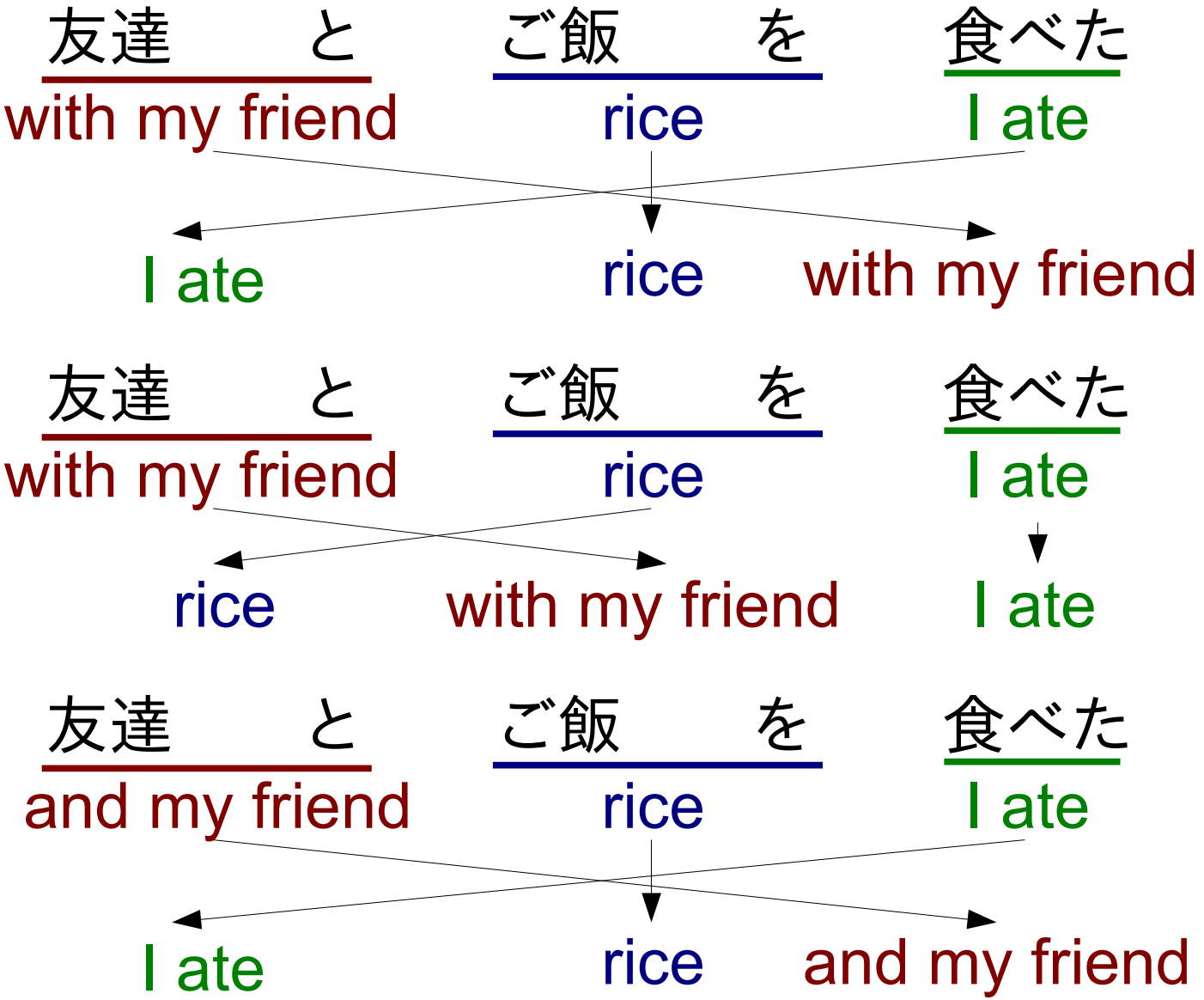


ランキング学習 [Hopkins+ 11]

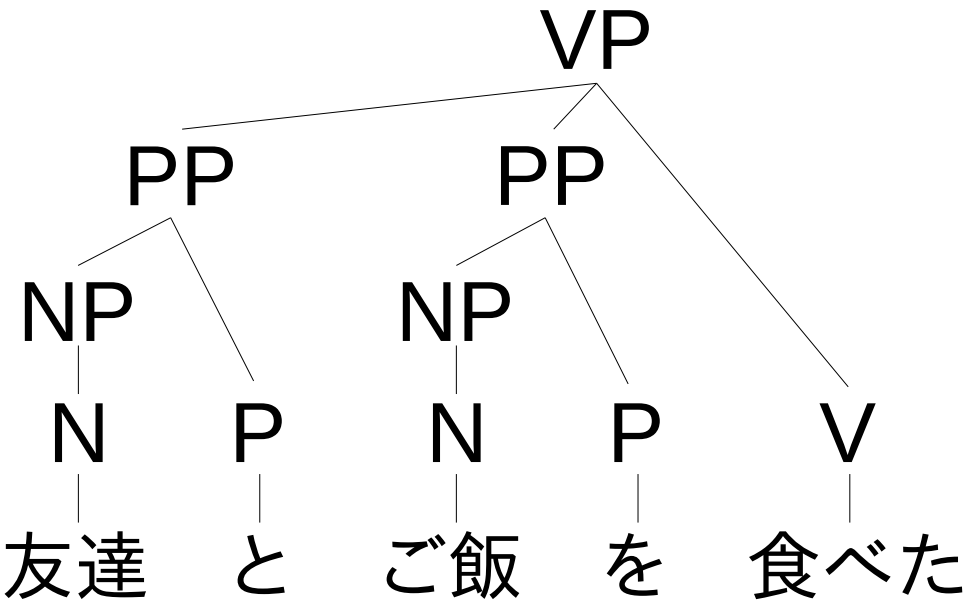
e1,1: #2
e1,2: #1
e1,3: #3

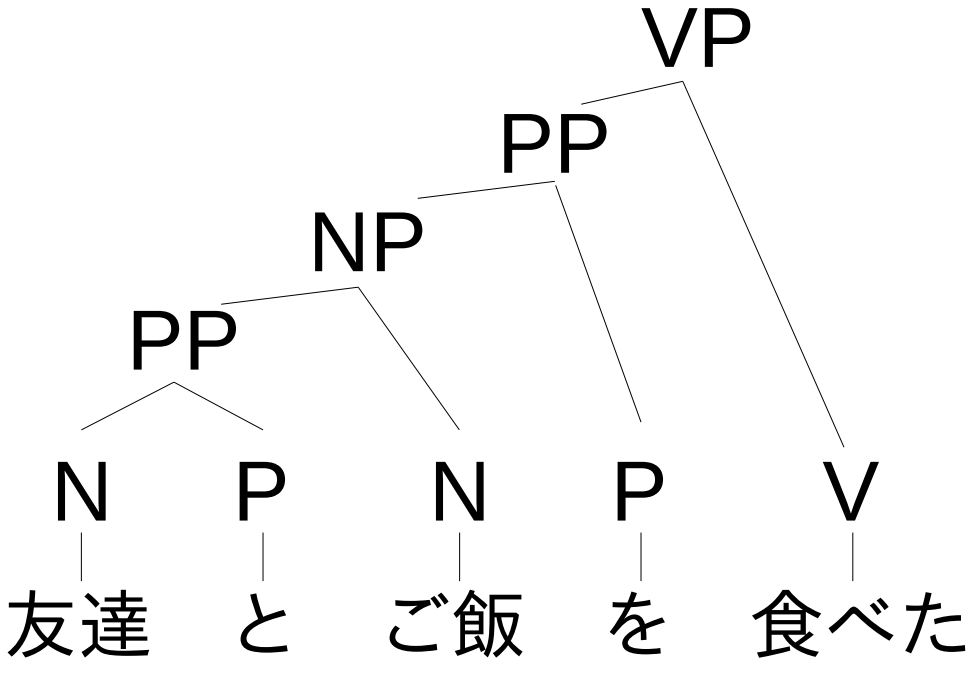
5. 統語ベース翻訳

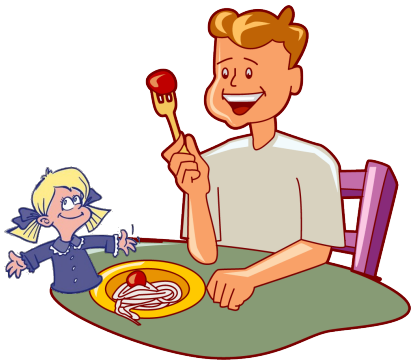
フレーズベース翻訳と並べ替え



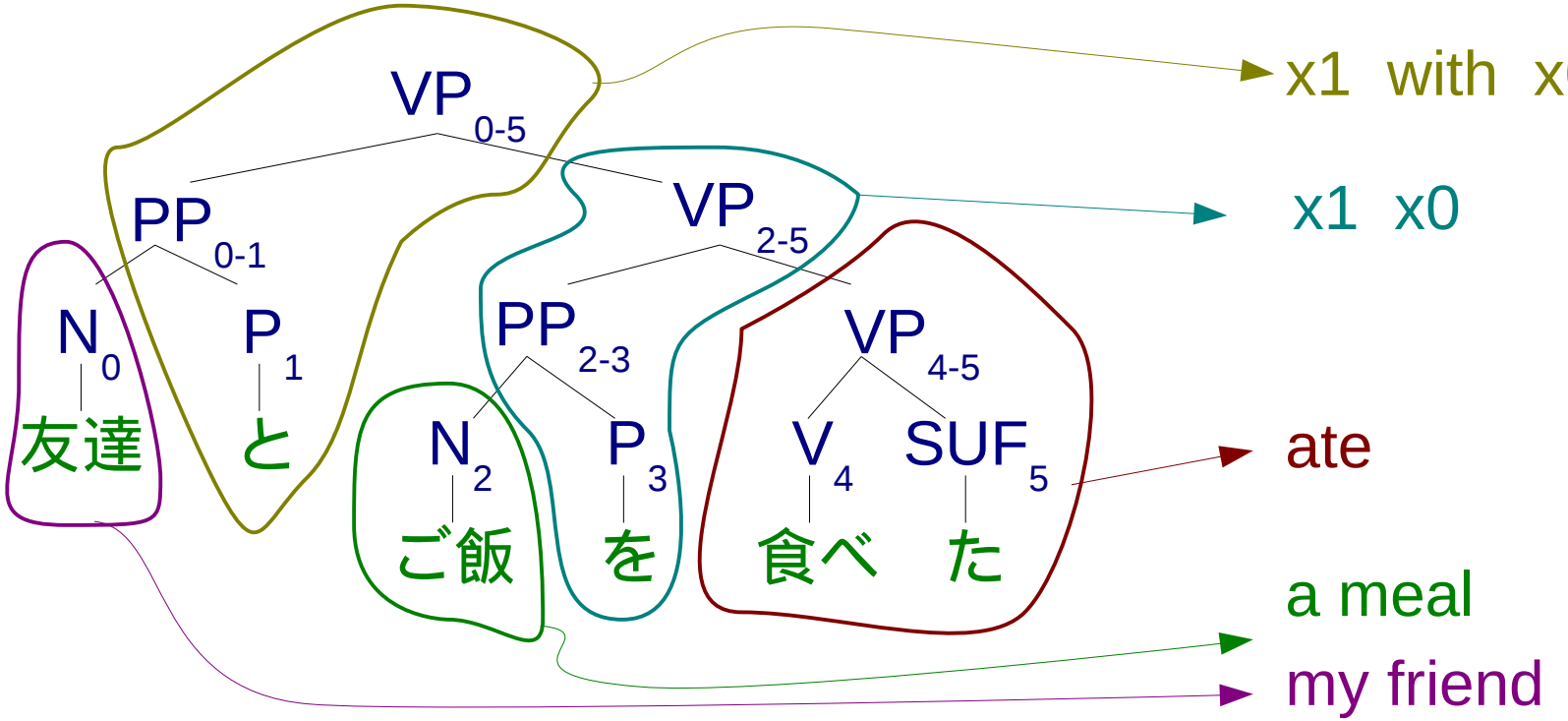
構文解析







Tree-to-String 翻訳 [Liu+ 06]



x1 x0

x1 x0

ate a meal with my friend

日本語における Tree-to-String 翻訳実験

- 少しの工夫で既存手法を大幅に上回る [Neubig+14]

入力	In equipment which generates CT by pressure waves, it was confirmed that development hours of CT bubbles increase in proportion to the 0.5th power of input power of pressure waves.
PB	気泡を生成することを確認した圧力波によりｃｔ装置において，０．５乗に比例して増加し，圧力波の入力電力のｃｔの開発時間。
T2S	圧力波によるｃｔを発生する装置において，ｃｔ気泡の開発時間が圧力波の入力パワーの０．５乗に比例することを確認した。
正解	圧力波によりｃｔを発生させる装置では圧力波の入力パワーの０．５乗に比例してｃｔ気泡の発達時間が増大することを確認した。

- オープンソース公開
  http://www.phontron.com/travatar

これからの機械翻訳？

文脈を考慮した機械翻訳

- 日英翻訳の主語の省略

あ、結構です。書かなくていいですよ。

正解： Oh, that's fine. You don't have to write.

システム：I don't write in, thank you.

- 文脈から読み取る語彙選択

Please keep lying.

正解： このまま休んでいてください。

システム：嘘をつけ続けてください。

翻訳の頑健性

- （Web 上などの）崩れた表現

はい、ごくろうさまでした。

正解：　You're all set.

システム：Yes, it was very good wax.

はい、　ごく　ろう　さま　でした　。

- 音声認識結果の入力等

同時性の高い音声翻訳
[Fujita+13, Oda+14]

同時性の高い音声翻訳
[Fujita+13, Oda+14]

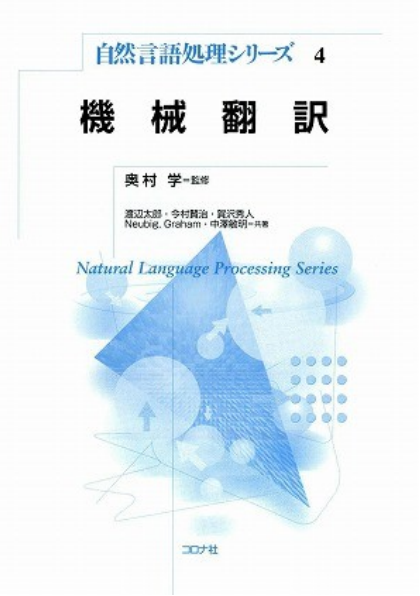
非言語情報を考慮した音声翻訳 [Kano+ 12, Kano+13]

非言語情報を考慮した音声翻訳 [Kano+ 12, Kano+13]

Thank you for listening!
↓
ご清聴ありがとうございます！

更に勉強するには：

コロナ社「機械翻訳」



ALAGIN セミナーの発表資料
https://sites.google.com/site/coronamachinetranslation/slides