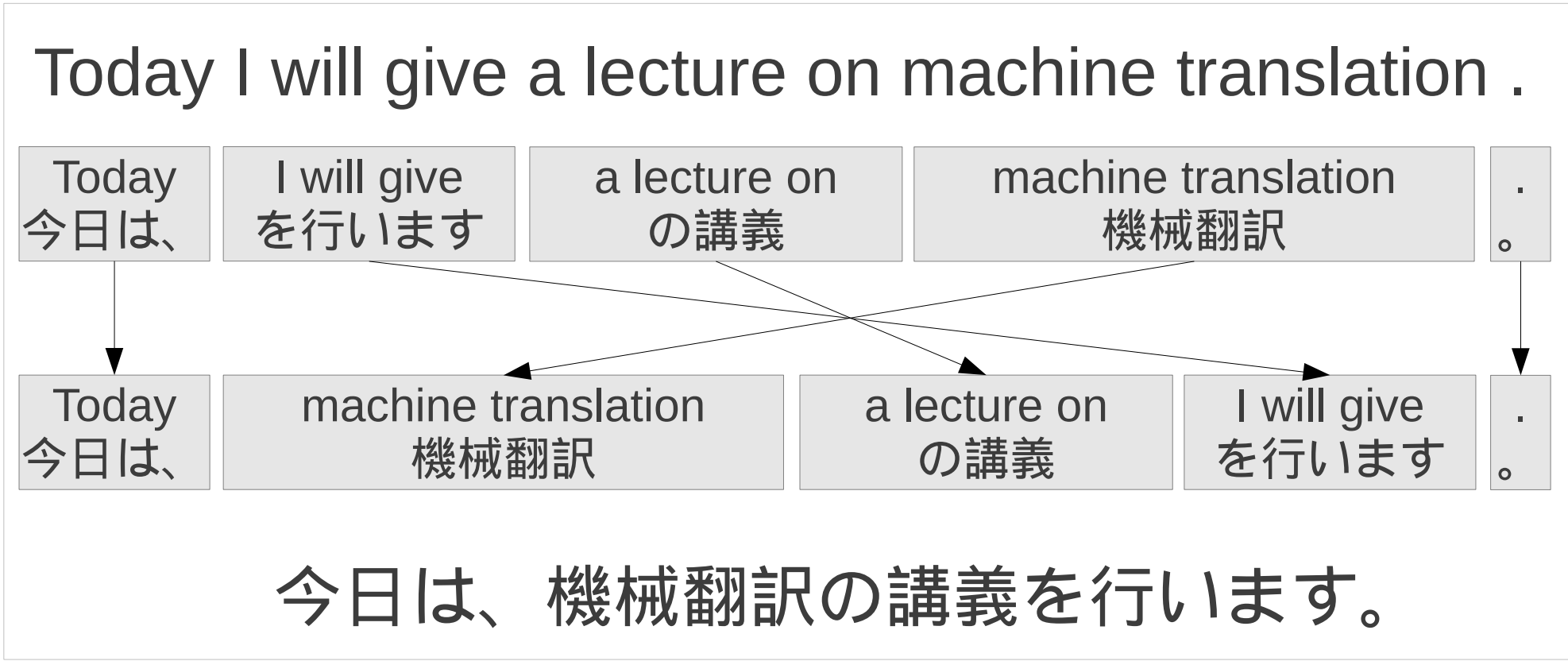
フレーズベース機械翻訳システムの構築

Graham Neubig & Kevin Duh
奈良先端科学技術大学院大学 (NAIST)
5/10/2012

フレーズベース 統計的機械翻訳（SMT）

- 文を翻訳可能な小さい塊に分けて、並べ替える



- 翻訳モデル・並べ替えモデル・言語モデルをテキストから統計的に学習

発表内容

1） フレーズベース統計的機械翻訳を構築する時に必要となる作業のステップ。

2） オープンソース機械翻訳システム Moses* の中で各ステップを実装したツール。

3） 各ステップにおける研究・未解決の問題。

* http://www.statmt.org/moses

フレーズベース統計的機械翻訳システムの構築の流れ

- データ収集
- トークン化
- 言語モデル
- アライメント
- フレーズ抽出 /Scoring
- Reordering Models
- 探索（デコーディング）
- 評価
- チューニング

データ収集

データ収集

- 文ごとの並列データ（パラレルデータ）
  - 翻訳モデル・並べ替えモデルの学習に利用

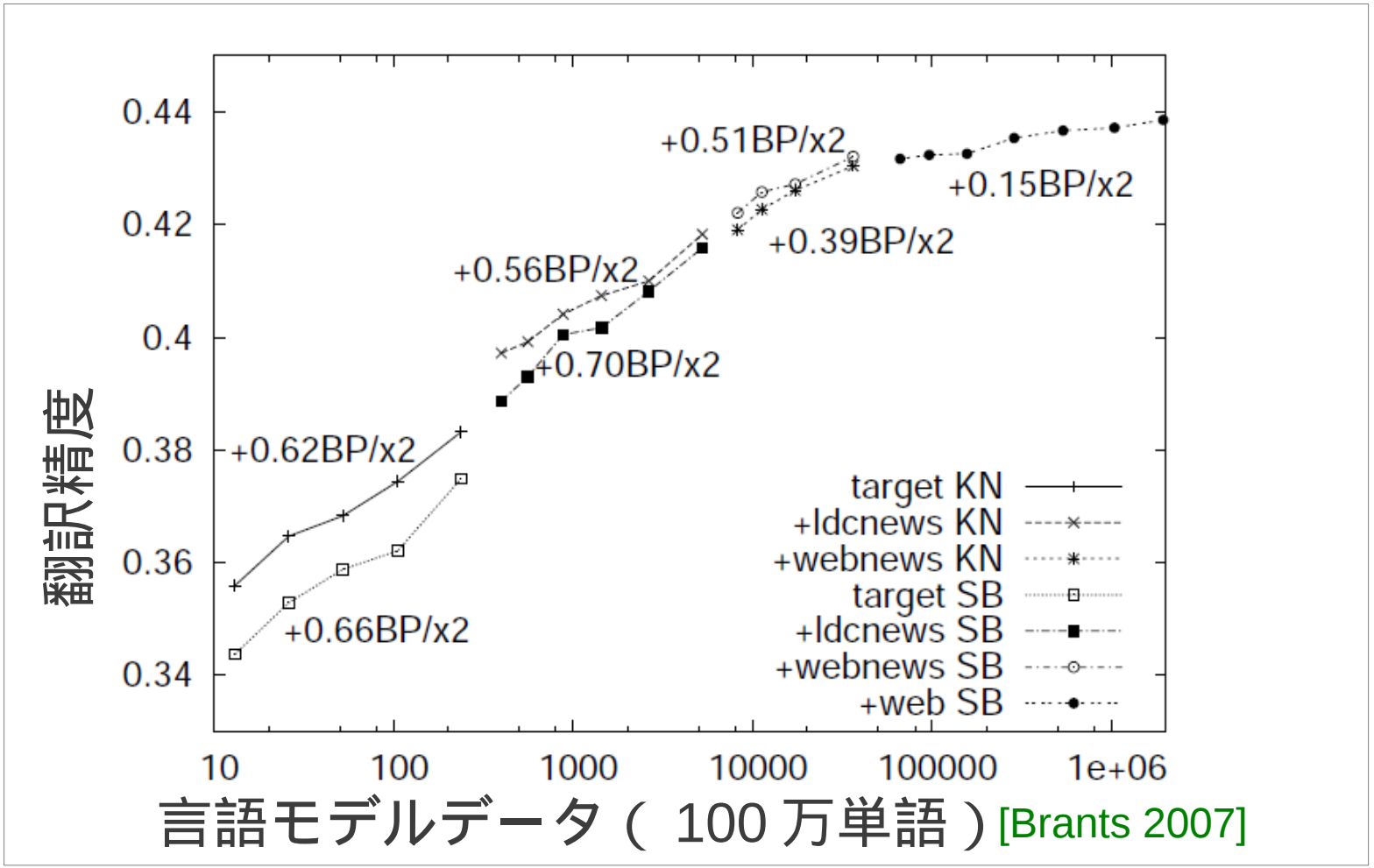
これはペンです。	This is a pen.
昨日は友達と食べた。	I ate with my friend yesterday.
象は鼻が長い。	Elephants' trunks are long.

- 単言語データ（目的言語側）
  - 言語モデルの学習に利用

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

翻訳に役立つデータは

- 大きい
→



[Brants 2007]

- 翻訳の質が高い、翻訳でない文を含まない
- テストデータと同一の分野

データ収集

- ワークショップ等では用意・指定されている

例:
IWSLT 2011 →

Name	Type	Words
TED	Lectures	1.76M
News Commentary	News	2.52M
EuroParl	Political	45.7M
UN	Political	301M
Giga	Web	576M

- 実用システムでは：
  - 各国政府・自治体・新聞
  - Web データ
  - 複数のデータ源の組み合わせ

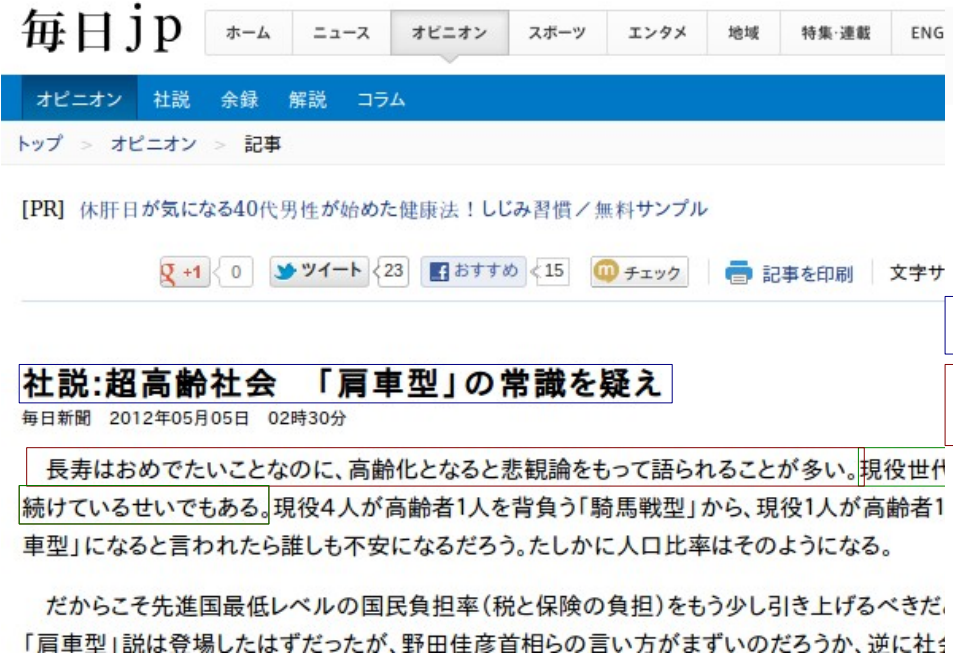
研究

- 並列ページの発見 [Resnik 03]

[画像： 毎日新聞]

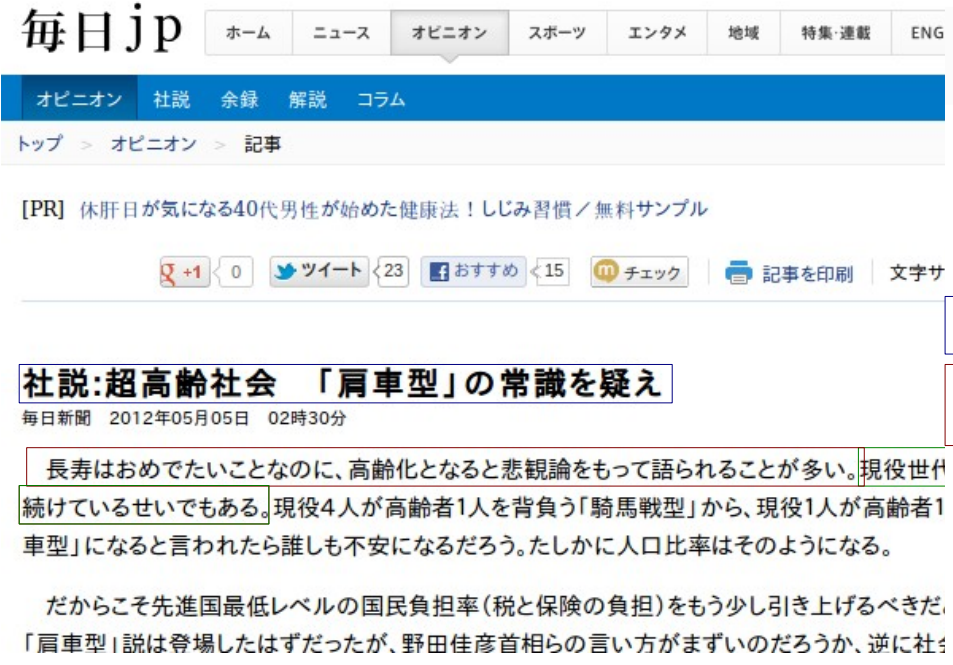
研究

- 並列ページの発見 [Resnik 03]
- 文アライメント [Moore 02]



研究

- 並列ページの発見 [Resnik 03]
- 文アライメント [Moore 02]



- データ作成のクラウドソーシング [Ambati 10]
  - Mechanical Turk 、 duolingo 等

トークン化

トークン化

- 例：日本語の単語分割

太郎が花子を訪問した。

↓

太郎 が 花子 を 訪問 した 。

- 例：英語の小文字化、句読点の分割

Taro visited Hanako.

↓

taro visited hanako .

トークン化ツール

- ヨーロッパの言語

```
tokenize.perl en < input.en > output.en
tokenize.perl fr < input.fr > output.fr
```

- 日本語

MeCab: `mecab -O wakati < input.ja > output.ja`
KyTea: `kytea -notags < input.ja > output.ja`
JUMAN, etc.

- 中国語

Stanford Segmenter, LDC, KyTea, etc...

研究

- 機械翻訳の精度向上につながるトークン化
  - 精度が重要か、一貫性が重要か [Chang 08]
  - 他の言語に合わせた単語挿入 [Sudoh 11]

太郎 が 花子 を 訪問 した 。

Taro <ARG1> visited <ARG2> Hanako .

- 活用の処理（韓国語、アラビア語等） [Niessen 01]

단어란 도대체 무엇일까요？

↓

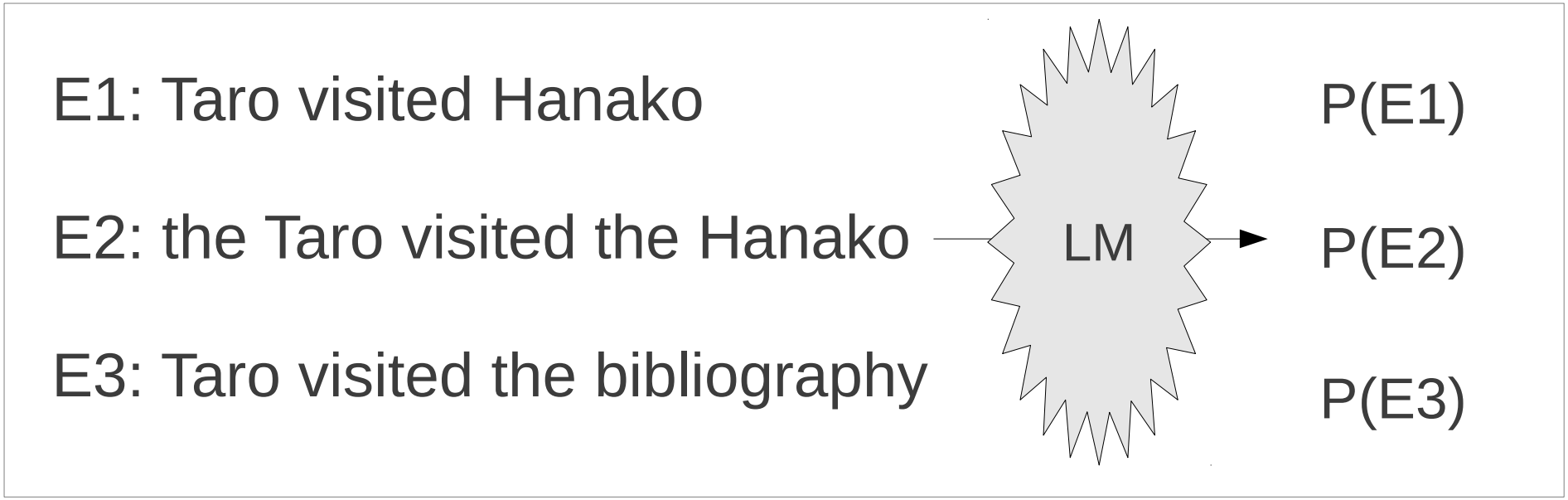
단어 란 도대체 무엇 일 까요 ？

- 教師なし学習 [Chung 09, Neubig 12]

言語モデル

言語モデル

- 目的言語側の各文に確率を与える



- 良い言語モデル：流暢性の高い文に高い確率を

P(E1) > P(E2)　　P(E1) > P(E3)

n-gram モデル

- 以下の文の確率を求めるとする

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

- n-gram モデル： 1 単語ずつ確率を計算
  - 直前の n-1 単語を考慮した条件付き確率
    例： 2-gram モデル

$$P(w_1=\text{“Taro”}) * P(w_2=\text{”visited”} \mid w_1=\text{“Taro”})$$
$$* P(w_3=\text{”Hanako”} \mid w_2=\text{”visited”})$$
$$* P(w_4=\text{”</s>”} \mid w_3=\text{”Hanako”})$$

注：
文末記号 </s>

ツール

- SRILM:

  学習：

```
ngram-count -order 5 -interpolate -kndiscount -unk
     -text input.txt -lm lm.arpa
```

  テスト：

```
ngram -lm lm.arpa -ppl test.txt
```

- ほかに： KenLM, RandLM, IRSTLM

研究

- n-gram に勝てるものはあるのか？ [Goodman 01]
  - 計算がシンプルで高速
  - 探索アルゴリズムと相性が良い
  - シンプルな割に強力
- その他の手法
  - 統語情報を利用した言語モデル [Charniak 03]
  - ニューラルネット言語モデル [Bengio 06]
  - モデル M [Chen 09]
  - などなど…

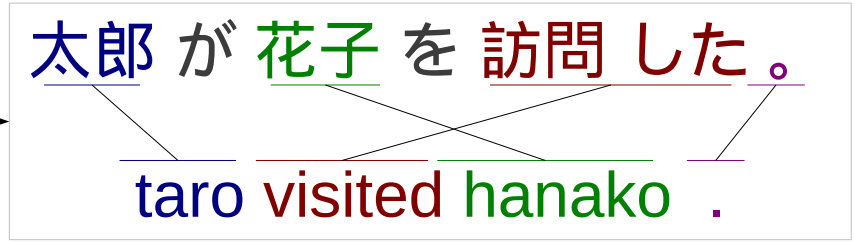
アライメント

アライメント

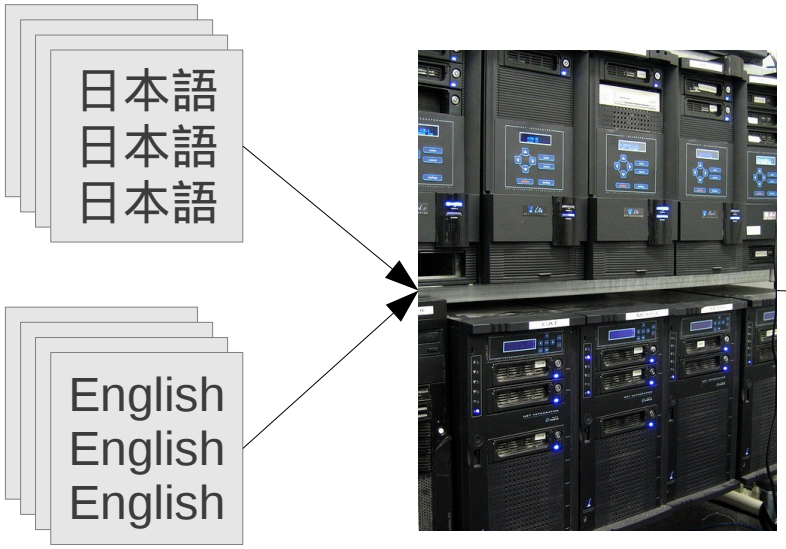
- 文内の単語対応を発見

太郎 が 花子 を 訪問 した 。

taro visited hanako .



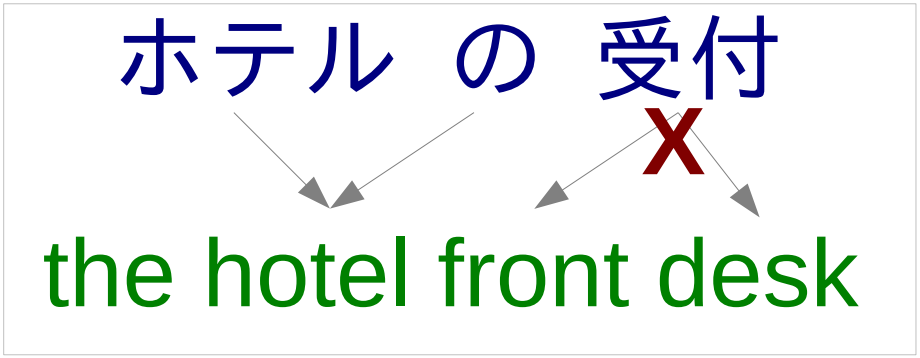
- 確率モデルによる自動学習（教師なし）が主流

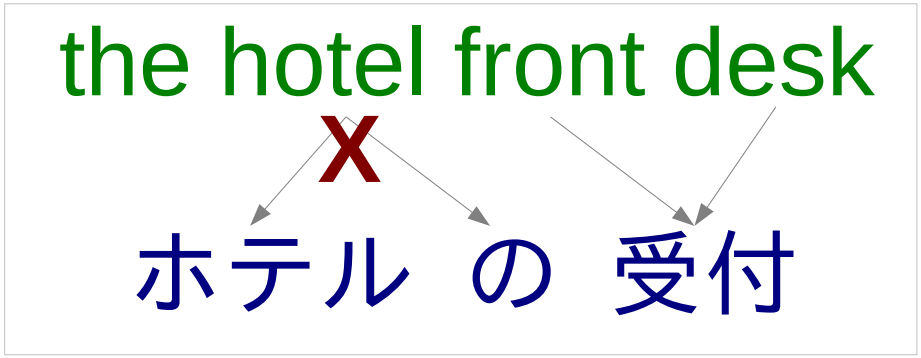


P(花子 |hanako) = 0.99
P(太郎 |taro) = 0.97
P(visited| 訪問) = 0.46
P(visited| した) = 0.04
P(花子 |taro) = 0.0001

IBM/HMM モデル

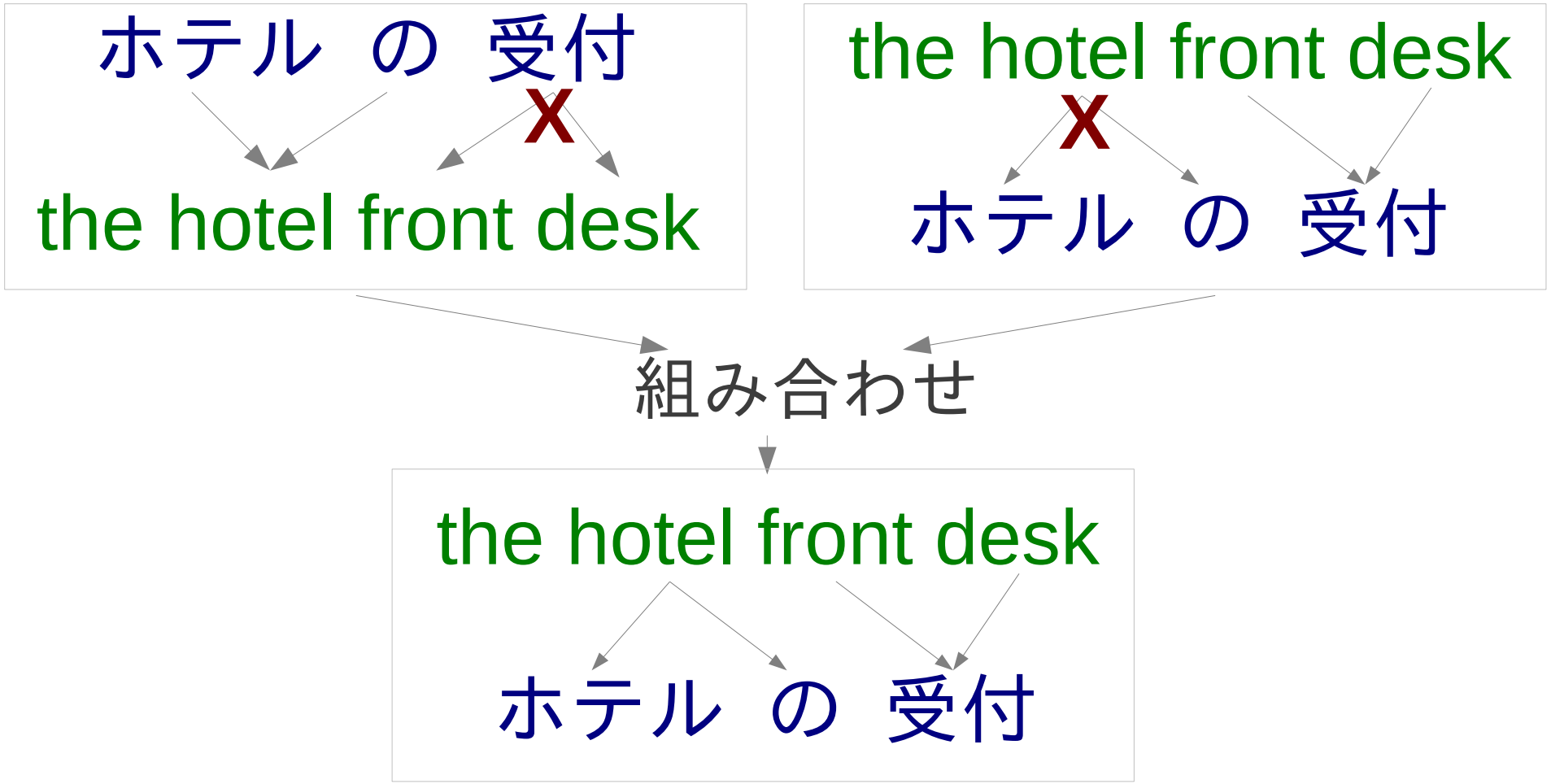
- １対多アライメントモデル





- IBM Model 1: 語順を考慮しない
- IBM Models 2-5, HMM: 徐々に考慮する情報を導入（精度・計算コスト ++ ）

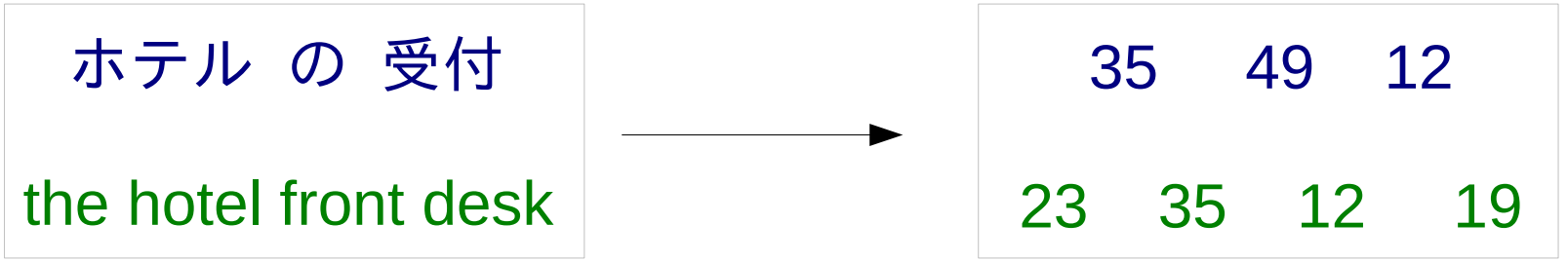
１対多アライメントの組み合わせ



- 様々なヒューリスティック手法（ grow-diag-final ）

ツール

- mkcls: 2 言語で単語クラスを自動発見



- GIZA++: IBM モデルによるアライメント（クラスを用いて確率を平滑化）

- symal: 両方向のアライメントを組み合わせる
- (Moses の train-model.perl の一部として実行される)

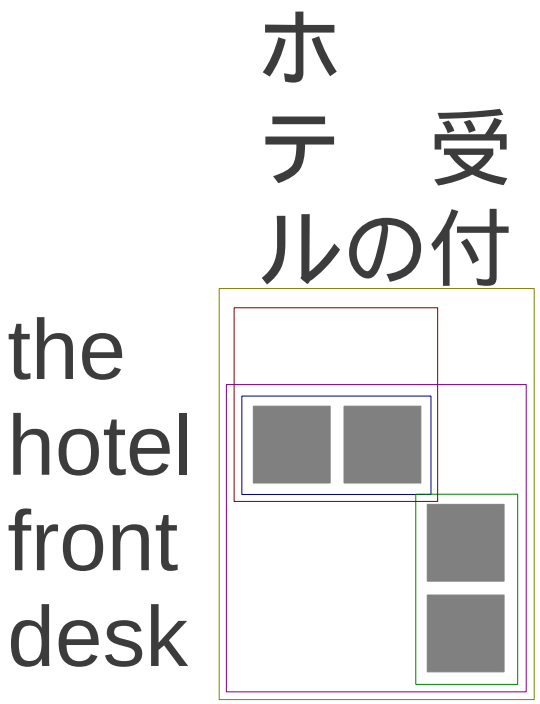
研究

- アライメントは本当に重要なのか？ [Aryan 06]
- 教師ありアライメント [Fraser 06, Haghighi 09]
- 統語情報を使ったアライメント [DeNero 07]
- フレーズベースアライメント [Marcu 02, DeNero 08]

フレーズ抽出

フレーズ抽出

- アライメントに基づいてフレーズを列挙



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

フレーズのスコア計算

- ５つの標準的なスコアでフレーズの信頼性・使用頻度
  - フレーズ翻訳確率
    $P(\mathbf{f}|\mathbf{e}) = c(\mathbf{f},\mathbf{e})/c(\mathbf{e})$　　$P(\mathbf{e}|\mathbf{f}) = c(\mathbf{f},\mathbf{e})/c(\mathbf{f})$

    例：　c(ホテル　の , the hotel) / c(the hotel)
  - 語彙 (lexical) 翻訳確率
    - フレーズ内の単語の翻訳確率を利用 (IBM Model 1)
    - 低頻度のフレーズ対の信頼度判定に役立つ

    $P(f|e) = \Pi_f 1/|\mathbf{e}| \sum_e P(f|e)$

    例：
    (P(ホテル |the)+P(ホテル |hotel))/2 * (P(の |the)+P(の |hotel))/2
  - フレーズペナルティ：すべてのフレーズで 1

ツール

- extract: フレーズ抽出
- phrase-extract/score: フレーズのスコア付け
- (Moses の train-model.perl の一部として実行される)

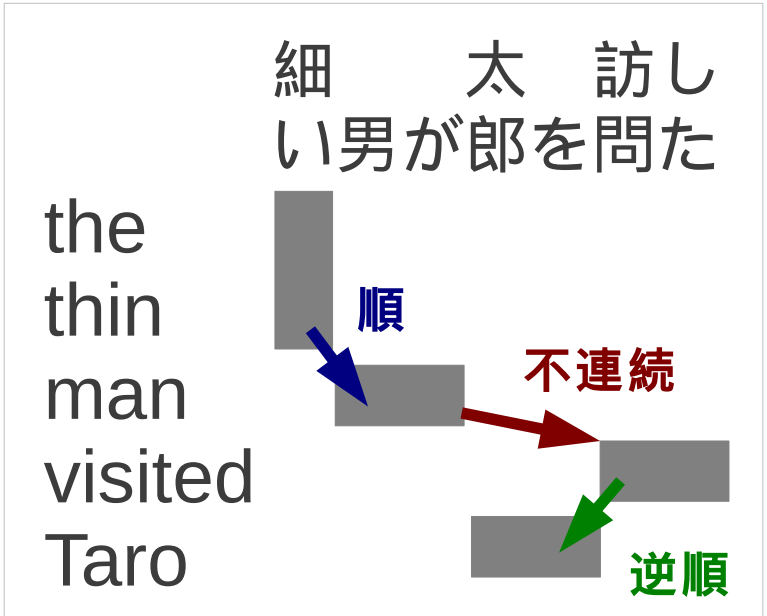
研究

- 翻訳モデルの分野適用 [Koehn 07, Matsoukas 09]
- 不要・信頼度の低いフレーズの削除 [Johnson 07]
- 一般化フレーズ抽出 (ソフト： Geppetto) [Ling 10]
- フレーズ曖昧性解消 [Carpuat 07]

並べ替えモデル

語彙化並べ替えモデル

- 順・逆順・不連続



細い → the thin
順の確率が高い

太郎 を → Taro
逆順の確率が高い

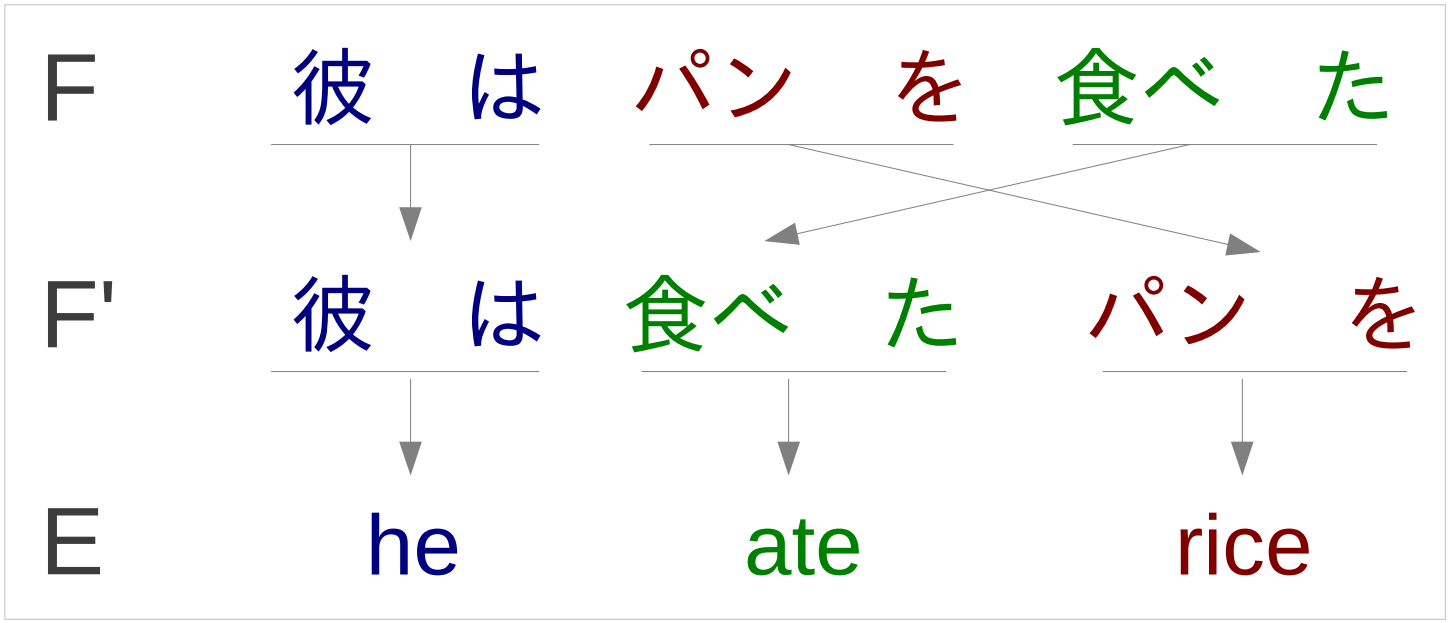
- 入力・出力、右・左などで条件付けた確率

ツール

- extract: フレーズ抽出と同一
- lexical-reordering/score: 並べ替えモデルを学習
- (Moses の train-model.perl の一部として実行される)

研究

- まだ未解決の問題が多い (特に日英・英日)
- 翻訳モデル自体の変更
  - 階層的フレーズベース翻訳 [Chiang 07]
  - 統語ベース翻訳 [Yamada 01, Galley 06]
- 前並べ替え [Xia 04, Isozaki 10]



探索
（デコーディング）

探索

- モデルによる最適な解を探索（または n-best ）

- 厳密な解を求めるのは NP 困難問題 [Knight 99]
- ビームサーチを用いて近似解を求める [Koehn 03]

ツール

- Moses!
  ```
  moses -f moses.ini < input.txt > output.txt
  ```

- その他 : moses_chart, cdec (階層的フレーズ、統語モデル)

研究

- レティス入力の探索 [Dyer 08]
- 統語ベース翻訳の探索 [Mi 08]
- 最小ベイズリスク [Kumar 04]
- 厳密な解の求め方 [Germann 01]

評価

人手評価

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

自動評価

- システム出力は正解文に一致するか
  - （翻訳の正解は単一ではないため、複数の正解も利用）
- BLEU: n-gram 適合率 + 短さペナルティ [Papineni 03]

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

- METEOR (類義語の正規化), TER (正解文に直すための変更数), RIBES (並べ替え)

研究

- 焦点を絞った評価尺度
  - 並べ替え [Isozaki 10]
  - 意味解析を用いた尺度 [Lo 11]
- チューニングに良い評価尺度 [Cer 10]
- 複数の評価尺度の利用 [Albrecht 07]
- 評価のクラウドソーシング [Callison-Burch 11]

チューニング

チューニング

- 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

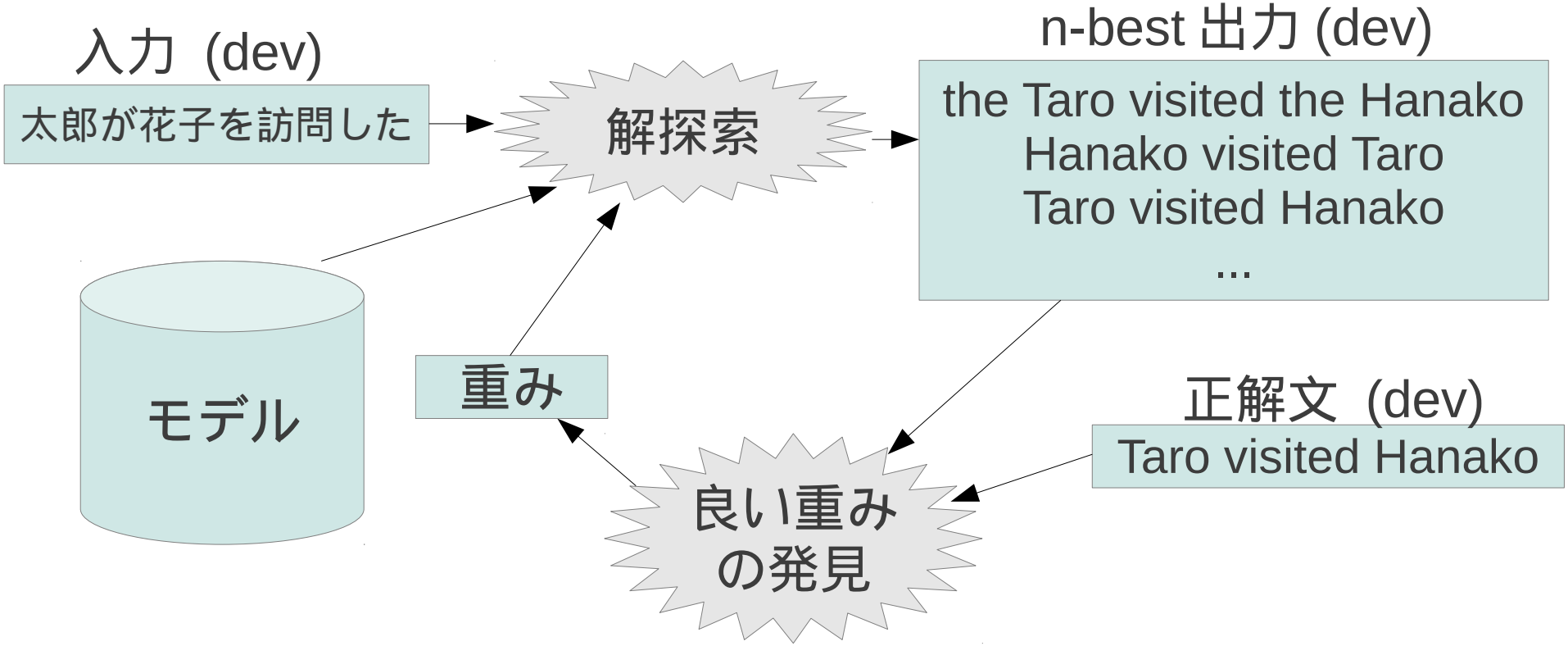
- スコアを重み付けると良い結果が得られる

	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

- チューニングは重みを発見 : $w_{LM}=0.2$ $w_{TM}=0.3$ $w_{RM}=0.5$

チューニング法

- 誤り最小化学習 : MERT [Och 03]



- その他 : MIRA [Watanabe 07] (オンライン学習), PRO (ランク学習) [Hopkins 11]

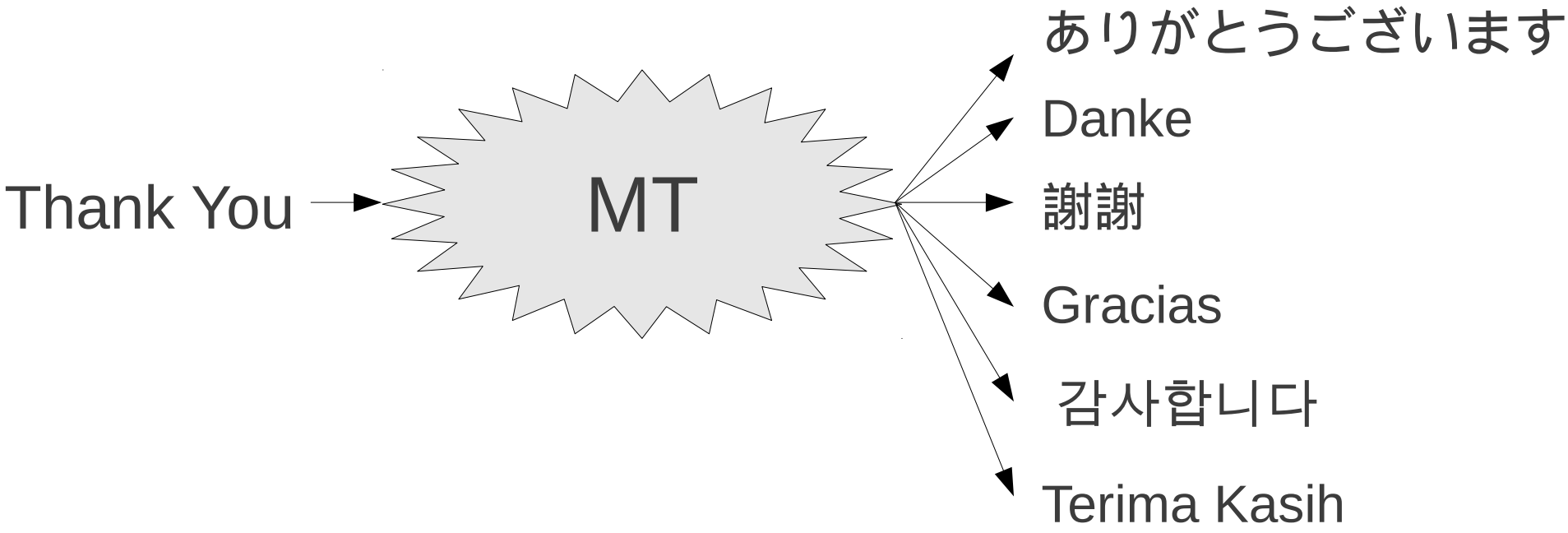
研究

- 膨大な素性数でチューニング (例 : MIRA, PRO)
- ラティス出力のチューニング [Macherey 08]
- チューニングの高速化 [Suzuki 11]
- 複数の評価尺度の同時チューニング [Duh 12]

おわりに

おわりに

- 機械翻訳は楽しい！一緒にやりましょう
- 年々精度が向上しているが、多くの問題が残る
- システムは大きいので、１つの部分に焦点を絞る



参考文献

- J. Albrecht and R. Hwa. A re-examination of machine learning approaches for sentence-level mt evaluation. In Proc. ACL, pages 880-887, 2007.
- V. Ambati, S. Vogel, and J. Carbonell. Active learning and crowdsourcing for machine translation. Proc. LREC, 7:2169-2174, 2010.
- N. Ayan and B. Dorr. Going beyond AER: an extensive analysis of word alignments and their impact on MT. In Proc. ACL, 2006.
- Y. Bengio, H. Schwenk, J.-S. Sencal, F. Morin, and J.-L. Gauvain. Neural probabilistic language models. In Innovations in Machine Learning, volume 194, pages 137-186. 2006.
- T. Brants, A. C. Popat, P. Xu, F. J. Och, and J. Dean. Large language models in machine translation. In Proc. EMNLP, pages 858-867, 2007.
- C. Callison-Burch, P. Koehn, C. Monz, and O. Zaidan. Findings of the 2011 workshop on statistical machine translation. In Proc. WMT, pages 22-64, 2011.
- M. Carpuat and D. Wu. How phrase sense disambiguation outperforms word sense disambiguation for statistical machine translation. In Proc. TMI, pages 43-52, 2007.
- D. Cer, C. Manning, and D. Jurafsky. The best lexical metric for phrasebased statistical MT system optimization. In NAACL HLT, 2010.
- P.-C. Chang, M. Galley, and C. D. Manning. Optimizing Chinese word segmentation for machine translation performance. In Proc. WMT, 2008.
- E. Charniak, K. Knight, and K. Yamada. Syntax-based language models for statistical machine translation. In MT Summit IX, pages 40-46, 2003.
- S. Chen. Shrinking exponential language models. In Proc. NAACL, pages 468-476, 2009.
- D. Chiang. Hierarchical phrase-based translation. Computational Linguistics, 33(2), 2007.
- T. Chung and D. Gildea. Unsupervised tokenization for machine translation. In Proc. EMNLP, 2009.
- J. DeNero, A. Bouchard-C^ote, and D. Klein. Sampling alignment structure under a Bayesian translation model. In Proc. EMNLP, 2008.
- J. DeNero and D. Klein. Tailoring word alignments to syntactic machine translation. In Proc. ACL, volume 45, 2007.
- K. Duh, K. Sudoh, X. Wu, H. Tsukada, and M. Nagata. Learning to translate with multiple objectives. In Proc. ACL, 2012.
- C. Dyer, S. Muresan, and P. Resnik. Generalizing word lattice translation. In Proc. ACL, 2008.

- A. Fraser and D. Marcu. Semi-supervised training for statistical word alignment. In Proc. ACL, pages 769-776, 2006.
- M. Galley, J. Graehl, K. Knight, D. Marcu, S. DeNeefe, W. Wang, and I. Thayer. Scalable inference and training of context-rich syntactic translation models. In Proc. ACL, pages 961-968, 2006.
- U. Germann, M. Jahr, K. Knight, D. Marcu, and K. Yamada. Fast decoding and optimal decoding for machine translation. In Proc. ACL, pages 228-235, 2001.
- J. T. Goodman. A bit of progress in language modeling. Computer Speech & Language, 15(4), 2001.
- A. Haghighi, J. Blitzer, J. DeNero, and D. Klein. Better word alignments with supervised ITG models. In Proc. ACL, 2009.
- M. Hopkins and J. May. Tuning as ranking. In Proc. EMNLP, 2011.
- H. Isozaki, T. Hirao, K. Duh, K. Sudoh, and H. Tsukada. Automatic evaluation of translation quality for distant language pairs. In Proc. EMNLP, pages 944-952, 2010.
- H. Isozaki, K. Sudoh, H. Tsukada, and K. Duh. Head nalization: A simple reordering rule for sov languages. In Proc. WMT and MetricsMATR, 2010.
- J. H. Johnson, J. Martin, G. Foster, and R. Kuhn. Improving translation quality by discarding most of the phrasetable. In Proc. EMNLP, pages 967-975, 2007.
- K. Knight. Decoding complexity in word-replacement translation models. Computational Linguistics, 25(4), 1999.
- P. Koehn, F. J. Och, and D. Marcu. Statistical phrase-based translation. In Proc. HLT, pages 48-54, 2003.
- P. Koehn and J. Schroeder. Experiments in domain adaptation for statistical machine translation. In Proc. WMT, 2007.
- S. Kumar and W. Byrne. Minimum bayes-risk decoding for statistical machine translation. In Proc. HLT, 2004.
- W. Ling, T. Lus, J. Graca, L. Coheur, and I. Trancoso. Towards a General and Extensible Phrase-Extraction Algorithm. In M. Federico, I. Lane, M. Paul, and F. Yvon, editors, Proc. IWSLT, pages 313-320, 2010.
- C.-k. Lo and D. Wu. Meant: An inexpensive, high-accuracy, semiautomatic metric for evaluating translation utility based on semantic roles. In Proc. ACL, pages 220-229, 2011.
- W. Macherey, F. Och, I. Thayer, and J. Uszkoreit. Lattice-based minimum error rate training for statistical machine translation. In Proc. EMNLP, 2008.
- D. Marcu and W. Wong. A phrase-based, joint probability model for statistical machine translation. In Proc. EMNLP, 2002.

- S. Matsoukas, A.-V. I. Rosti, and B. Zhang. Discriminative corpus weight estimation for machine translation. In Proc. EMNLP, pages 708717, 2009.
- H. Mi, L. Huang, and Q. Liu. Forest-based translation. In Proc. ACL, pages 192-199, 2008.
- R. Moore. Fast and accurate sentence alignment of bilingual corpora. Machine Translation: From Research to Real Users, pages 135-144, 2002.
- G. Neubig, T. Watanabe, S. Mori, and T. Kawahara. Machine translation without words through substring alignment. In Proc. ACL, Jeju, Korea, 2012.
- S. Niessen, H. Ney, et al. Morpho-syntactic analysis for reordering in statistical machine translation. In Proc. MT Summit, 2001.
- F. J. Och. Minimum error rate training in statistical machine translation. In Proc. ACL, 2003.
- K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. BLEU: a method for automatic evaluation of machine translation. In Proc. COLING, pages 311-318, 2002.
- P. Resnik and N. A. Smith. The web as a parallel corpus. Computational Linguistics, 29(3):349-380, 2003.
- J. Suzuki, K. Duh, and M. Nagata. Distributed minimum error rate training of smt using particle swarm optimization. In Proc. IJCNLP, pages 649-657, 2011.
- T. Watanabe, J. Suzuki, H. Tsukada, and H. Isozaki. Online largemargin training for statistical machine translation. In Proc. EMNLP, pages 764-773, 2007.
- F. Xia and M. McCord. Improving a statistical MT system with automatically learned rewrite patterns. In Proc. COLING, 2004.
- K. Yamada and K. Knight. A syntax-based statistical translation model. In Proc. ACL, 2001.
- O. F. Zaidan and C. Callison-Burch. Crowdsourcing translation: Professional quality from non-professionals. In Proc. ACL, pages 1220-1229, 2011.