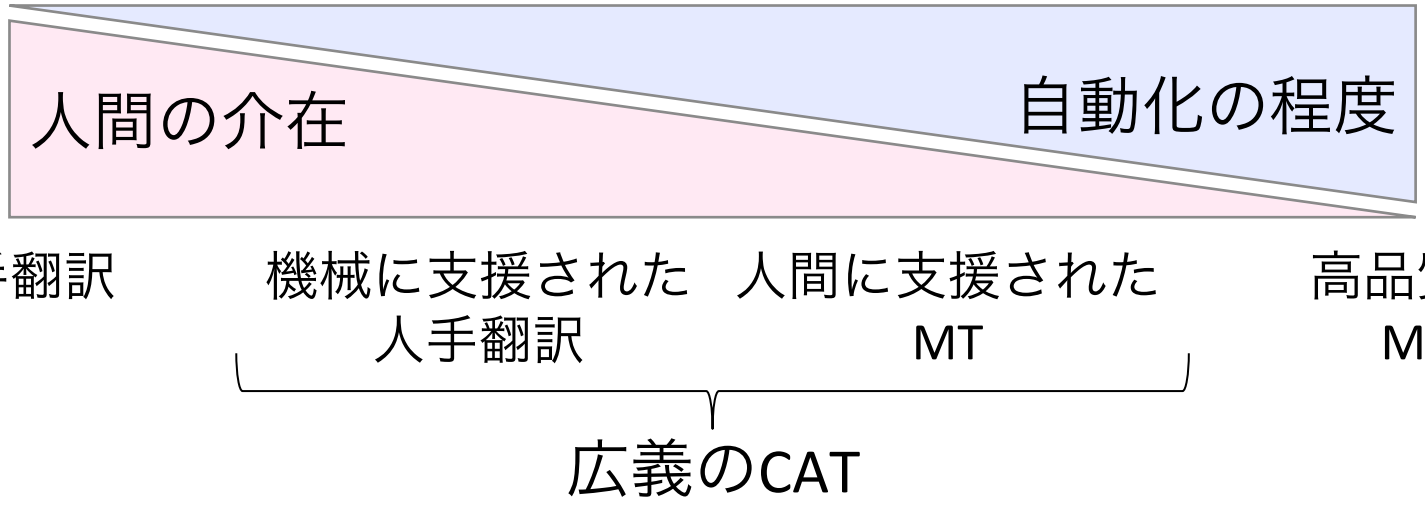
統計的機械翻訳とニューラル機械翻訳の混合nベストリランキング

Benjamin MARIE　　藤田 篤
情報通信研究機構 (NICT)
先進的音声翻訳研究開発推進センター (ASTREC)

NLPテーマセッションでの議論

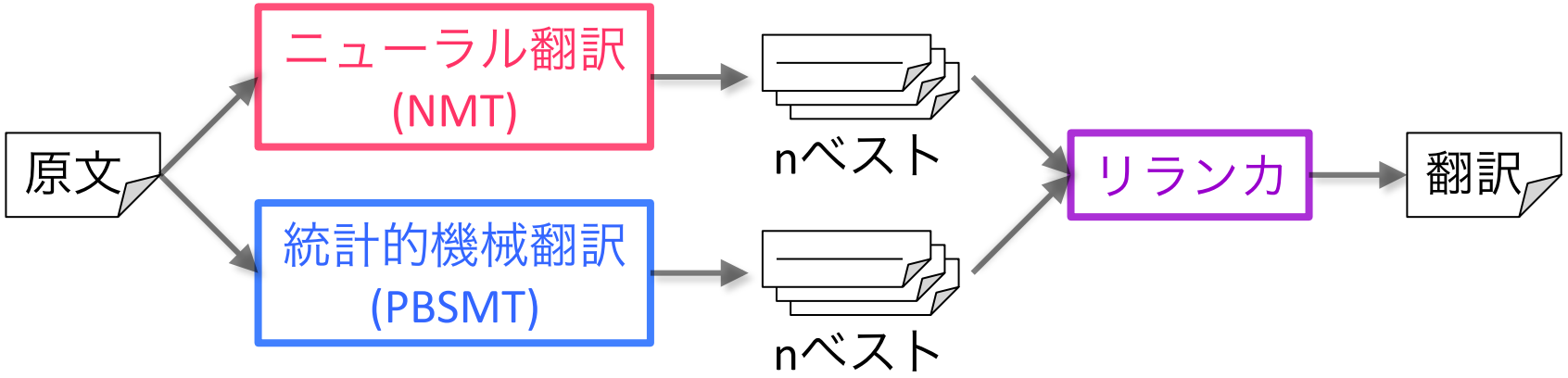
- 前回: 品質と効率，既存の翻訳戦略 [Hutchinson+, 92]



- 今回: 人間と機械の協働(CAT)に着目
  - 本発表: 後編集(PE)の必要性を認めた上でのMTの改善
    - MT単体は完璧ではない → 少しでも良くしたい
    - PE向け下訳の自動生成 → 多少遅くてもOK
      - cf. 日常会話の音声翻訳
    - F/W: 実際のPE負荷の調査 [山田+, 18b]

PBSMTとNMTってどっちが性能が良い?

- 両者を組み合わせるのが良い
  - Keywords: system combination, hybrid

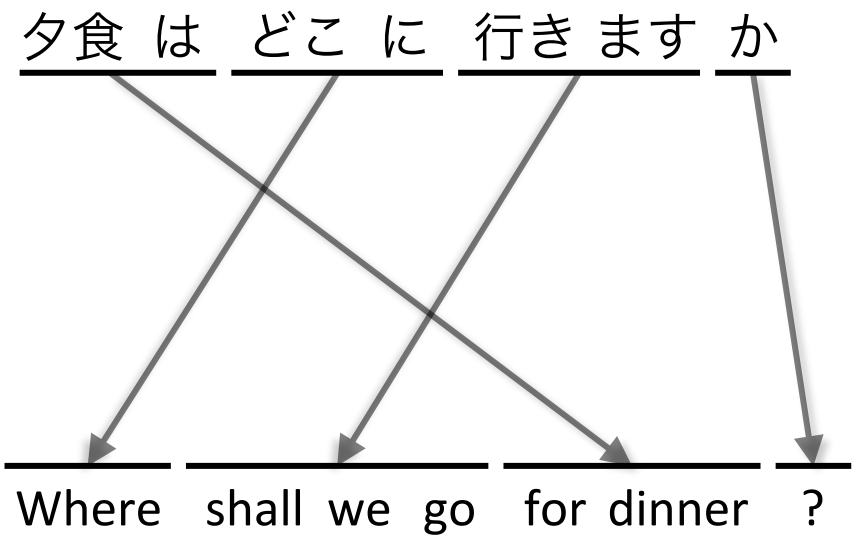


- リランキング用に種々の素性を検討
- 4つのタスクで一貫した性能向上を確認
  - 日⇔英: 特許翻訳
  - 仏⇔英: ニュース翻訳

背景と動機

PBSMT

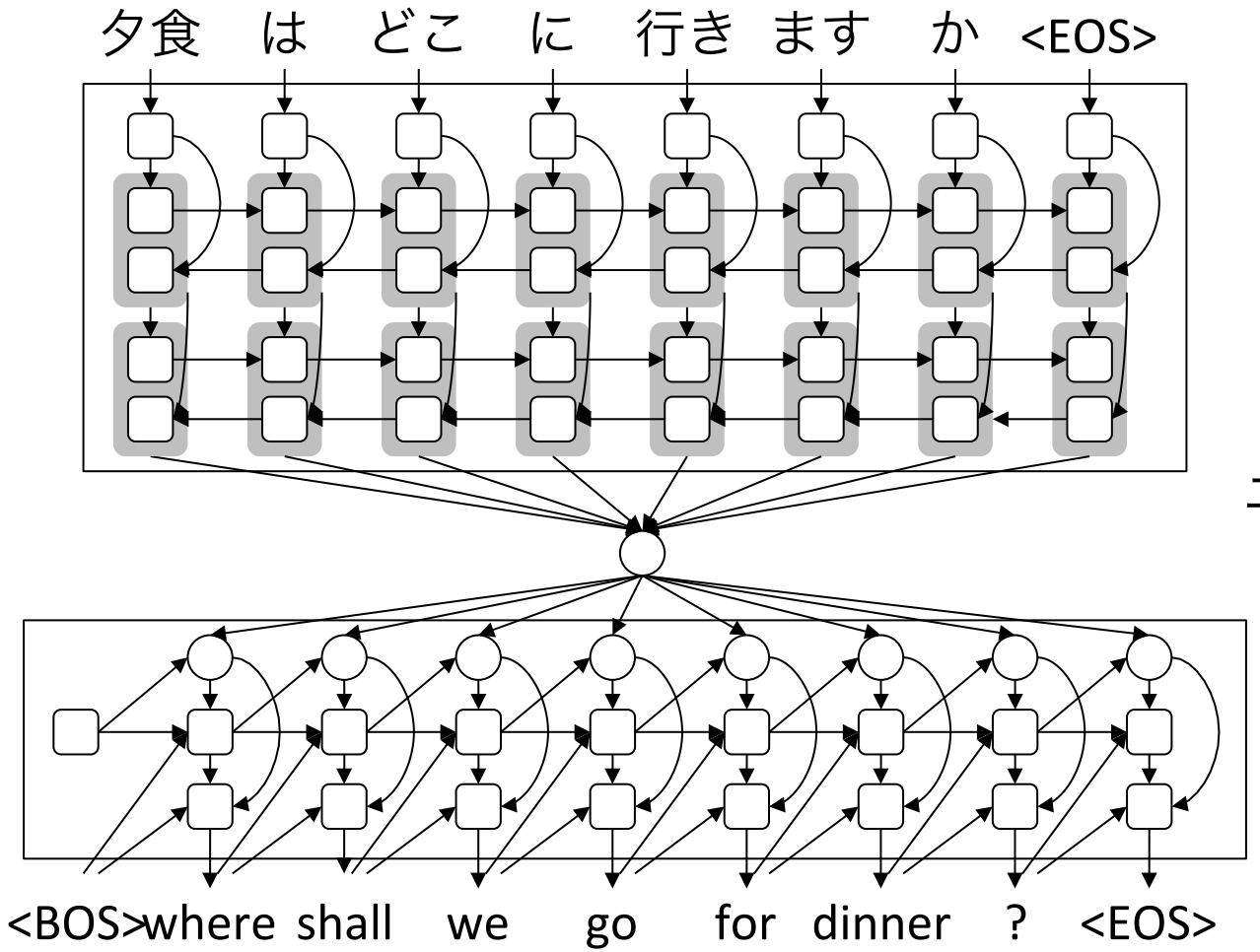
- 膨大な種類のパーツの組み合わせから最適解を探索
  - 原文中の様々な句の対訳候補の列挙 [Koehn+, 03]
  - 様々な並び替えの可能性の列挙
  - 流暢さの評価



対応関係を
陽に捉える

NMT

- 原文と各単語をベクトル空間で表現し，
  そこから目標言語の単語の系列を生成 [Bahdanau+, 14]

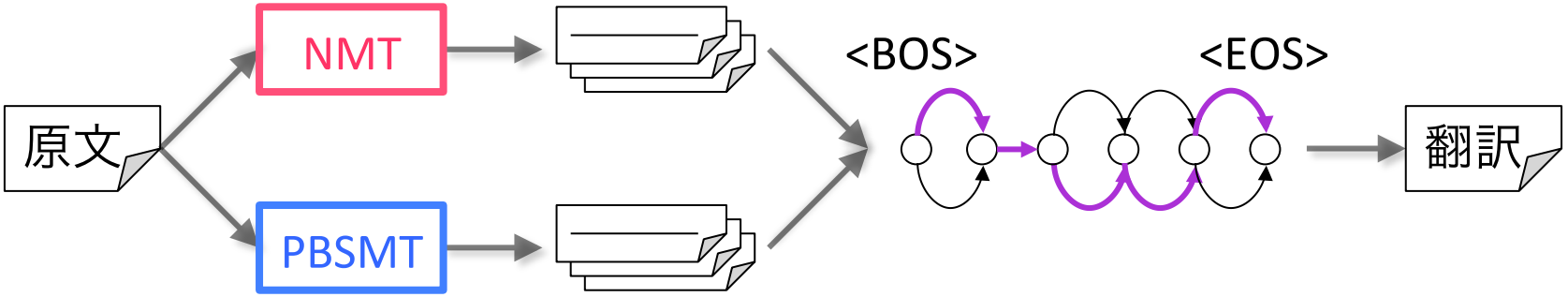


PBSMTとNMTってどっちが性能が良い?

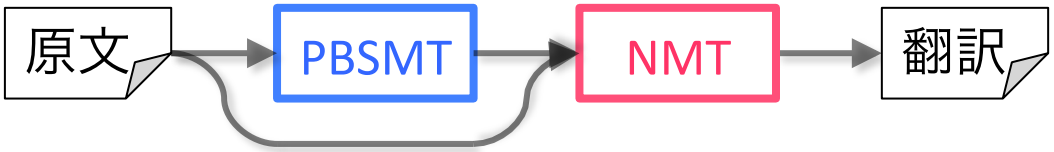
- 状況次第: 対訳データの有無，計算機資源など
- PBSMTはほぼ成熟済
  - 事前並べ替え [Isozaki+, 10][Goto+, 15]
  - 各種ニューラル素性 [Devlin+, 14][Mino+, 15]
  - 対訳データが小規模にしかない場合はNMTよりもマシ
- NMTは発展途上
  - 大規模データがあればPBSMTを上回る場合が多い
    - とくに流暢さの点で優秀
  - 課題
    - 遅い，GPUが不可欠
    - 重大な問題: 訳抜け，冗長な出力，類義語

ハイブリッドMT (1/2)

- 合議翻訳 [Bangalore+, 01][Watanabe+, 11][Freitag+, 14]
  - nベストにとらわれず，良いとこ取りができる可能性
  - 性能の劣化が生じる場合がある

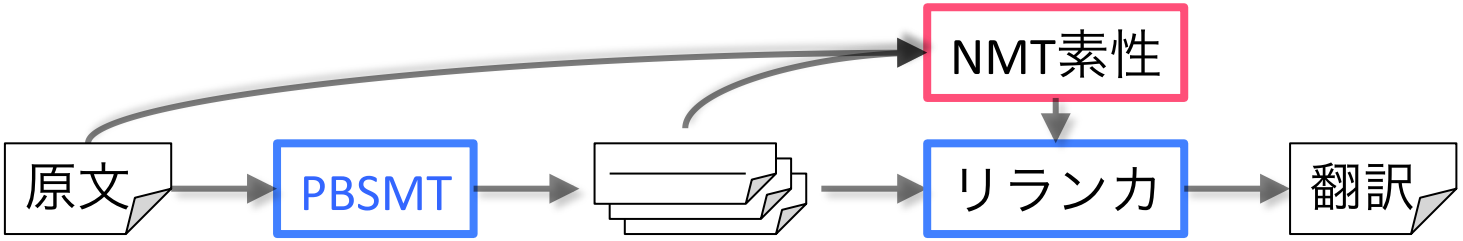


- 事前翻訳 (pre-translation) [Niehues+, 16]
  - 解ける部分を順番に解いていく
  - 実際にはうまくいかないケースが多い

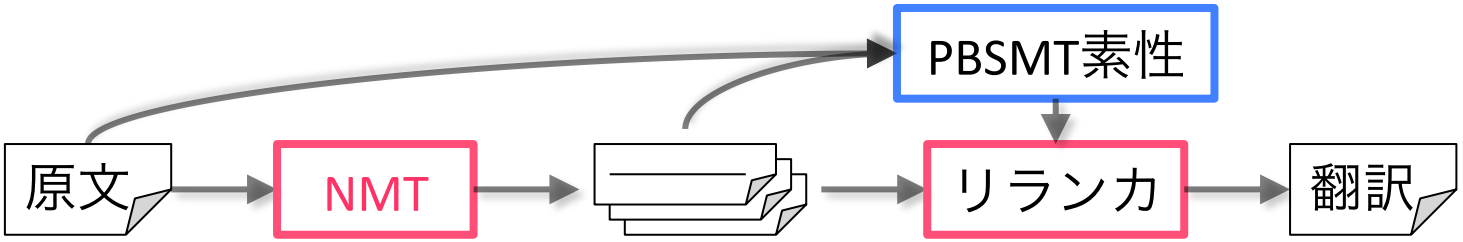


ハイブリッドMT (2/2)

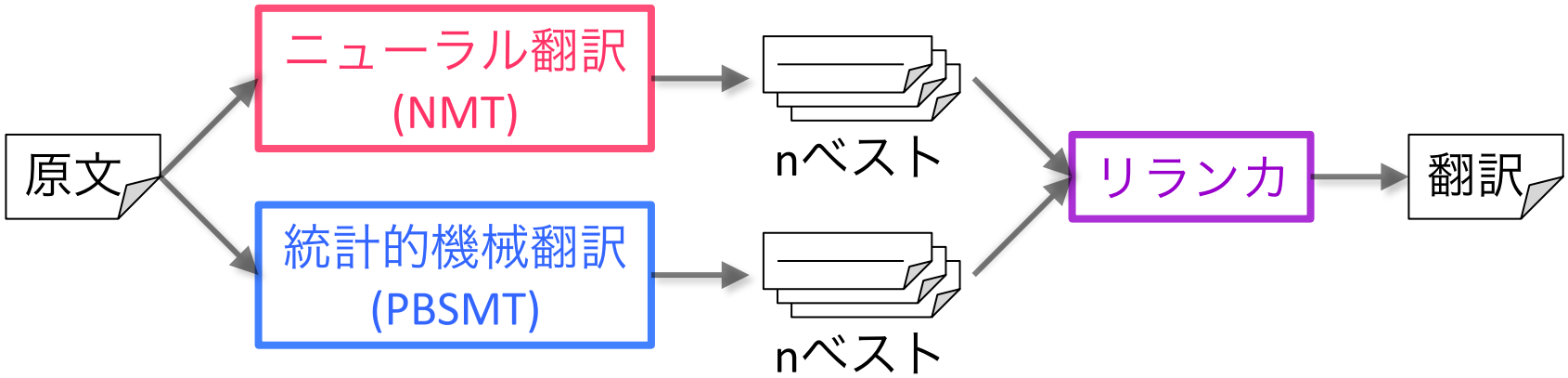
- nベスト候補のリランキング
  - PBSMTの出力 + NMTの尤度等 [Le+, 12][Sennrich+, 17]



  - NMTの出力 + PBSMTの尤度等 [Zhang+, 17]



提案手法: 混合nベストリランキング (良いとこ取り)



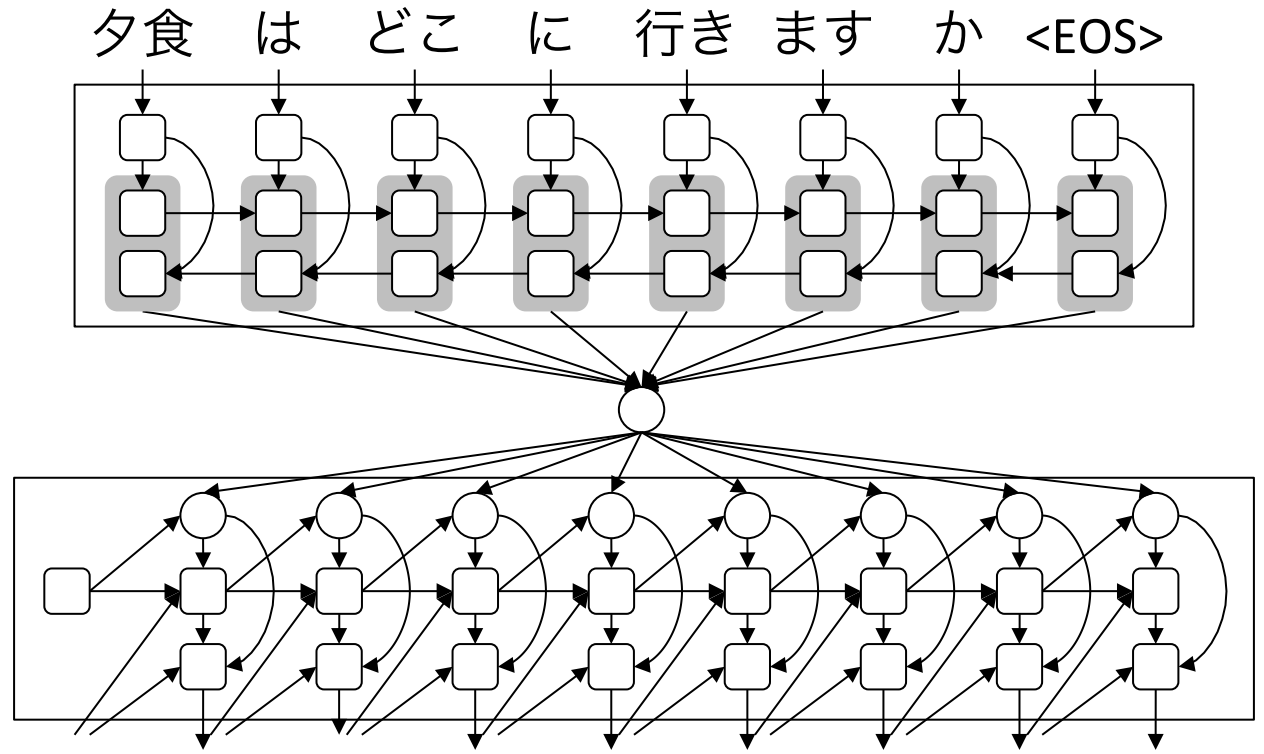
リランカの構築

- PBSMTのチューニング(MERT)と同じ [Och+, 03]
  - 各翻訳候補に素性を付与
  - 訓練: 参照訳を使って機械学習で重みをチューニング
  - 運用: 学習結果を使って各翻訳候補をリスコアリング

- 検討して最終的に残した素性
  - NMTに基づく素性
  - PBSMTに基づく素性
  - 語彙翻訳確率
  - 翻訳候補集合全体に基づくスコア
  - その他

NMTに基づく素性 (10種類)

- **L2R**: 訳文をleft-to-rightで生成
  - 4つのサブモデル，それらの幾何平均
- **R2L**: 訳文をright-to-leftで生成 [Sennrich+, 17]
  - 4つのサブモデル，それらの幾何平均



PBSMTに基づく素性 (1種類)

- **PBFD**: Phrase-based forced decoding [Zhang+, 17]
  - PBSMTのフレーズテーブルを使ってベストな対応を発見
  - 候補数膨大: 載っている対 + 単語の削除/挿入ルール

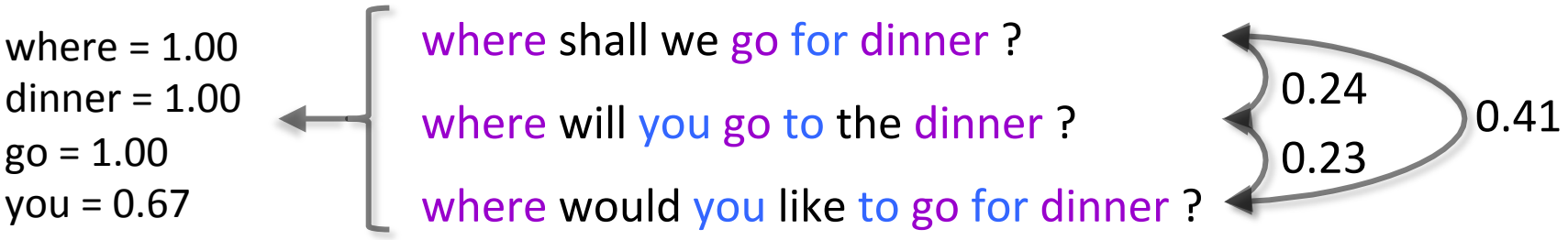
LEX: 語彙翻訳確率 (4種類)

- 対訳コーパスから単語アラインメントによって推定
  - 起点言語→目標言語: $P(e|j)$
  - 目標言語→起点言語: $P(j|e)$
- 文向けの合成法
  - 全組み合わせの平均 [Tillman+, 09]
  - 条件側の最大値の平均 [Hildebrand+, 08]

$P(e\|j)$	夕食	は	どこ	に	行きます	か
where	0.00	0.01	**0.47**	0.01	0.01	0.01	0.01
shall	0.00	**0.01**	0.00	0.01	0.01	0.00	0.01
we	0.00	0.01	0.01	0.01	**0.02**	0.01	0.01
go	0.00	0.01	0.02	0.01	**0.25**	0.01	0.01
for	0.00	**0.02**	0.00	0.02	0.01	0.01	0.01
dinner	**0.74**	0.00	0.00	0.00	0.00	0.00	0.00
?	0.00	0.01	0.06	0.01	0.01	0.02	**0.35**

翻訳候補集合全体に基づくスコア

- **WPP**: 単語事後確率 (2種類)
  - 多くの翻訳候補に出現する語を信頼する [Ueffing+, 07]
  - 元の文の重み付け
    - 語彙翻訳確率
    - NMTの尤度(L2R)
- **MBR**: コンセンサススコア (2種類)
  - 他の翻訳候補との平均的な類似度
    - 文レベルBLEU [Papineni+, 02]
    - chrF++ [Popović, 17]



その他の素性

- **LM**: 言語モデルのスコア (2種類)
- **WP**: 単語ペナルティ (1種類)
- **LEN** (2種類)
  - 原文と翻訳候補の長さの差分
  - その絶対値
- **SYS** (1種類)
  - NMTによる翻訳候補ならば1, そうでなければ0
    - TODO: NMTとPBSMTが同じ翻訳候補を出力する場合を要考慮

評価実験

十分なデータがある設定: 4言語方向

タスクとコーパス

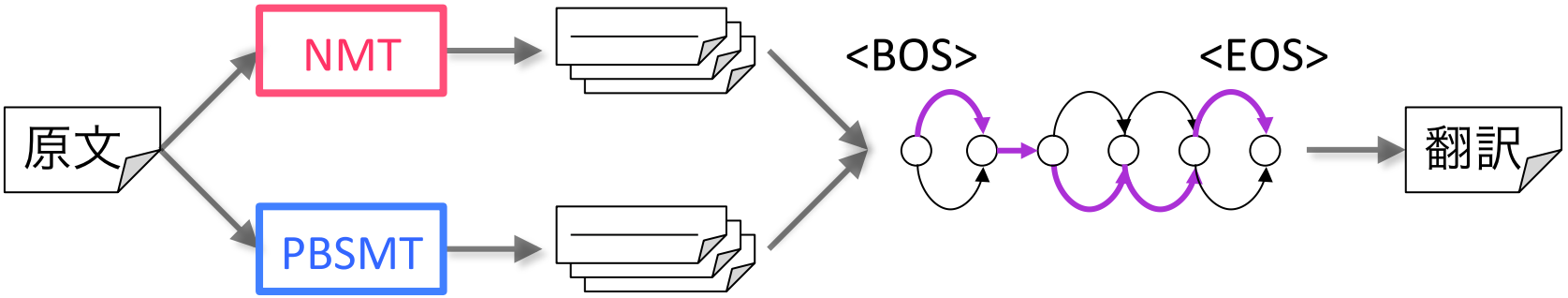
- 日⇔英: 特許翻訳 (NTCIR) [Goto+, 10]
  - 対訳
    - 訓練: 320万文
    - 開発: 2000文
    - 評価 (日英): 2000文 + 2300文
    - 評価 (英日): 2000文 + 2300文
  - 単言語: NTCIR (日: 270億トークン, 英: 150億トークン)
- 仏⇔英: ニュース翻訳 (WMT) [Bojar+, 15]
  - 対訳
    - 訓練: 2360万文
    - 開発: 3003文
    - 評価 (両方向): 3000文 + 3003文
  - 単言語: News Crawl (仏: 20億トークン, 英: 30億トークン)

翻訳システム

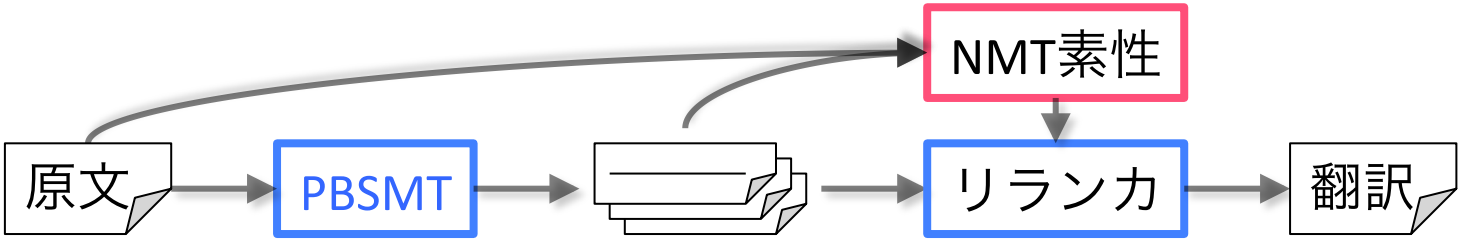
- PBSMT: Moses [Koehn+, 07]
  - 最長7語, grow-diag-final-and
  - 語彙化並び替えモデル, 双方向MSD
  - 4-gram言語モデル
    - 小: 訓練データの目標言語側
    - 大: 訓練データの目標言語側 + 単言語データ
  - Distortion-limit: 開発データを使ってチューニング, MIRA
- NMT: Nematus [Sennrich+, 17]
  - BPE適用後, 語彙サイズを5万に限定
  - モデル: 埋め込み層512次元, 隠れ層1000次元, 注意機構あり
  - 4つのL2Rモデルのアンサンブル
  - ビーム幅: 100, 文長によるスコアの正規化

リランキングシステム (既存手法)

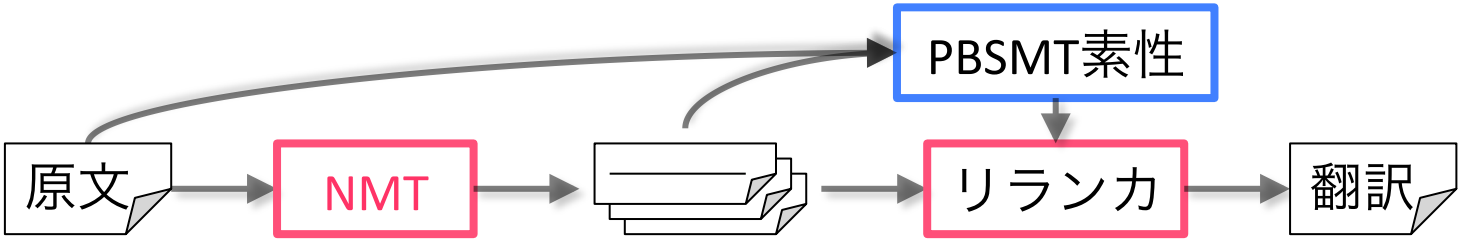
(a) 合議翻訳: Jane [Freitag+, 14] + PBSMTのLM



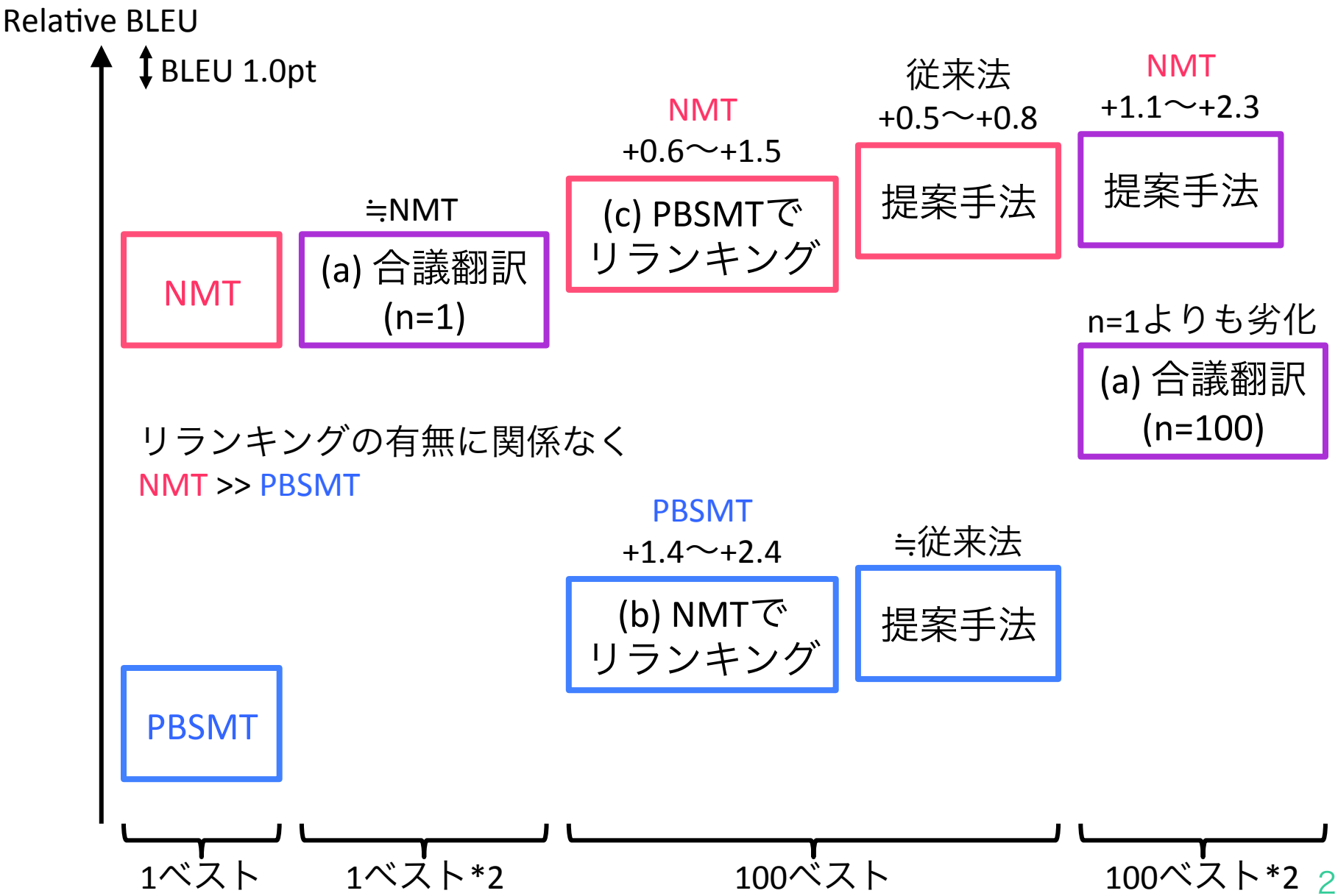
(b) PBSMTのnベスト + NMTの尤度 [Sennrich+, 17]



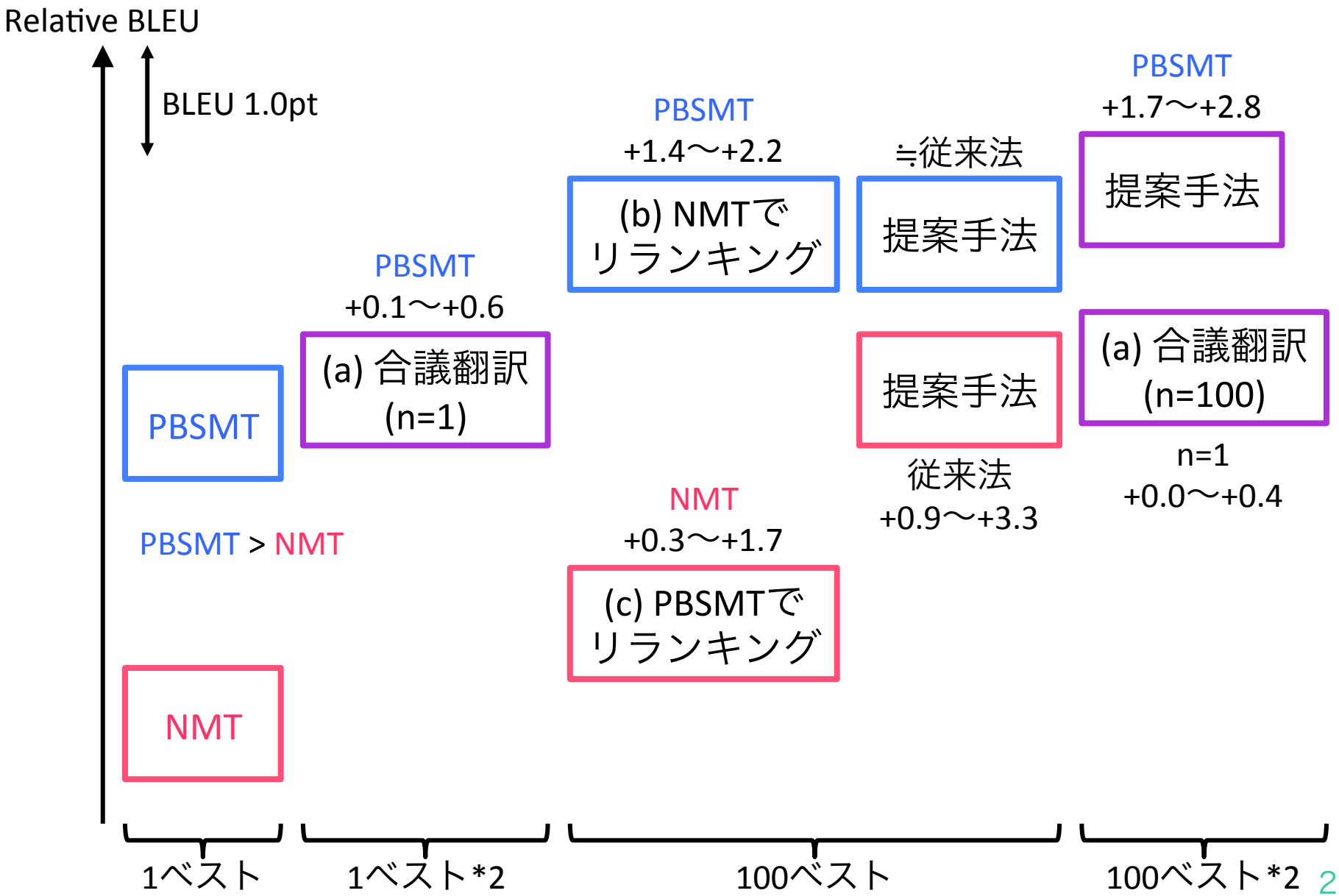
(c) NMTのnベスト + PBSMTの尤度 [Zhang+, 17]



手法間の優劣 (日英・英日特許翻訳)



手法間の優劣 (仏英・英仏ニュース翻訳)



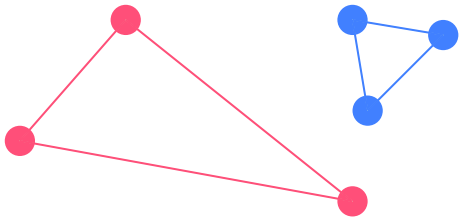
各素性の重要度

- 一部の素性を除外した実験に基づく考察
  - 1つ除外しても劣化の程度は小さい
    → 素性どうしがある程度相関していた

大分類	ラベル	素性数	説明	BLEUの変化と考察
NMT	**L2R**	5	翻訳確率(left-to-right)	-0.1〜0: 重要でない
NMT	**R2L**	5	翻訳確率(right-to-left)	-0.5〜-0.4: 常に重要
PBSMT	**PBFD**	1	翻訳確率(フレーズ)	-0.3〜+0.2: 意外と影響小
PBSMT	**LEX**	4	翻訳確率(語)	-0.2〜+0.1
候補集合	**WPP**	2	単語事後確率	-0.2〜+0.1
候補集合	**MBR**	2	コンセンサススコア	-0.3〜+0.1
その他	**LM**	2	言語モデル	-0.4〜+0.1: 仏英・英仏で重要
その他	**WP**	1	単語ペナルティ	-0.1〜+0.1
その他	**LEN**	2	長さの差と絶対値	-1.5〜+0.1: 仏英・英仏で重要
その他	**SYS**	1	候補の出自	-0.1〜+0.1

成功の理由: 候補の多様性

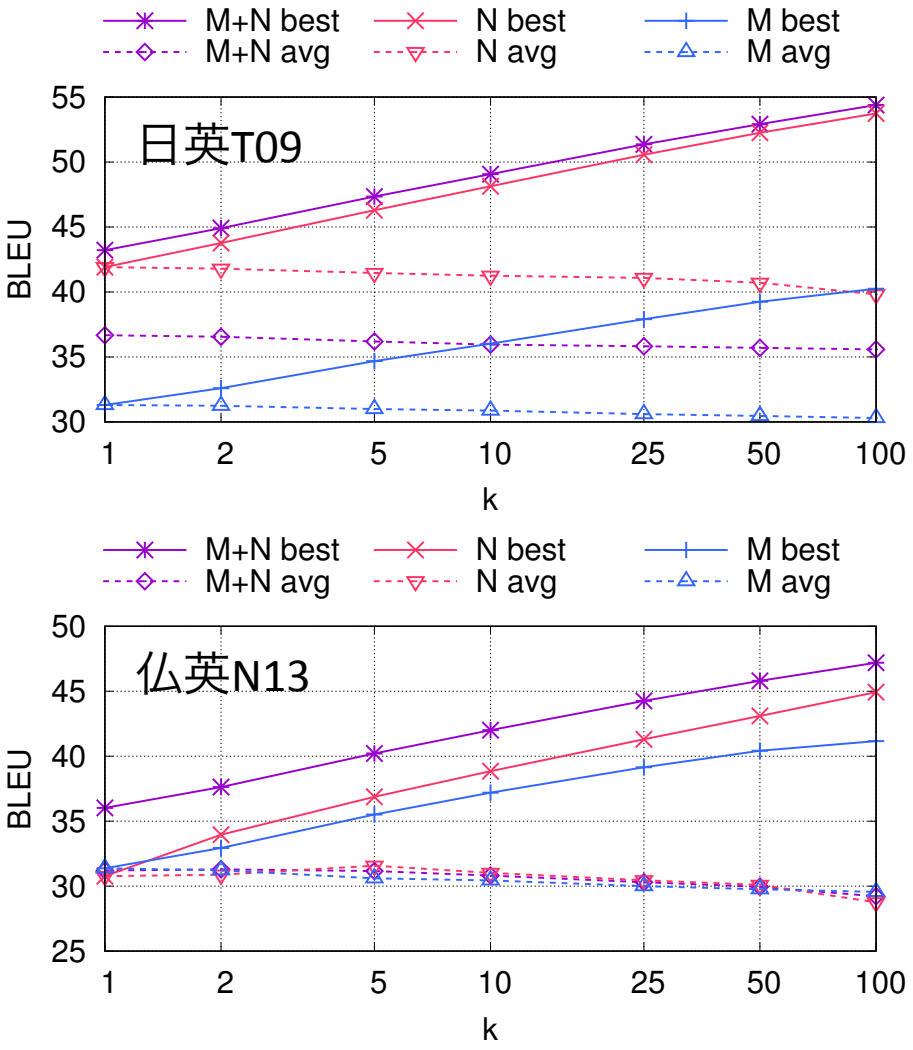
- リランキングでは候補の多様性が鍵 [Gimpel+, 13]
  - 2倍の種類の候補を参照
  - 候補間の平均的な類似度 (文レベルBLEU, chrF++)
    - (PBSMT+NMT) << NMT < PBSMT … 悪い候補も含まれうるが



  - 候補内のトークンの種類数
    - (PBSMT+NMT) > NMT > PBSMT … 和集合なので当然

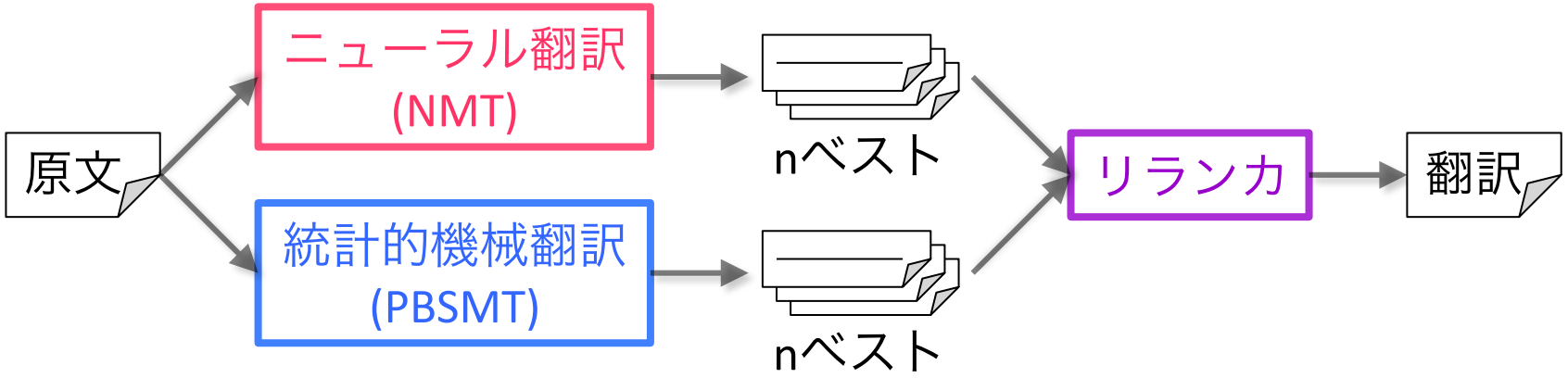
成功の理由: 候補の平均的な品質

- 各システムのnベストが割と安定して高品質
  - 日英・英日
    - Avg.: NMT >> PBSMT
    - Best: PBSMTを加える効果は小さそう
    - 実際: NMT +1.1〜+2.3
  - 仏英・英仏
    - Avg.: PBSMT ≒ NMT
    - Best: 混ぜると大幅改善
    - 実際: PBSMT +1.7〜+2.8
- F/W: サンプリング [Zhang+, 17][今村+, 18a]



まとめ

- 性質が異なる翻訳システムを組み合わせる
  - Keywords: system combination, hybrid



- リランキング用に種々の素性を検討
  - **R2L**が重要，**PBFD**は影響小，仏⇔英は**LEN** (文長差)が大事
- 4つのタスクで一貫した性能向上を確認
  - 日⇔英: 特許翻訳
  - 仏⇔英: ニュース翻訳

今後の課題

- 少資源な状況下での性能の検証
  - 小規模対訳データ
  - 分野適応
- State-of-the-artへの挑戦
  - より高精度な要素システムの利用
    - PBSMT: 事前並び替え，HIERO
    - NMT: 多層化，CNNエンコーダ
  - より多くのシステムの混合
  - サンプリング
- 外的評価
  - e.g., 下訳としての有用さ