Task-Oriented Word Segmentation

Tatsuya Hiraoka
博士論文
東京工業大学（岡崎研）
※現在 富士通株式会社

どんな話？

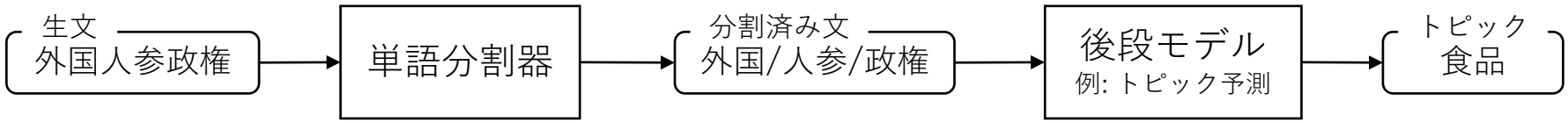
単語分割を**タスク**に応じて**最適化**する話

Task-Oriented Word Segmentation

Tatsuya Hiraoka
博士論文
東京工業大学（岡崎研）
※現在 富士通株式会社

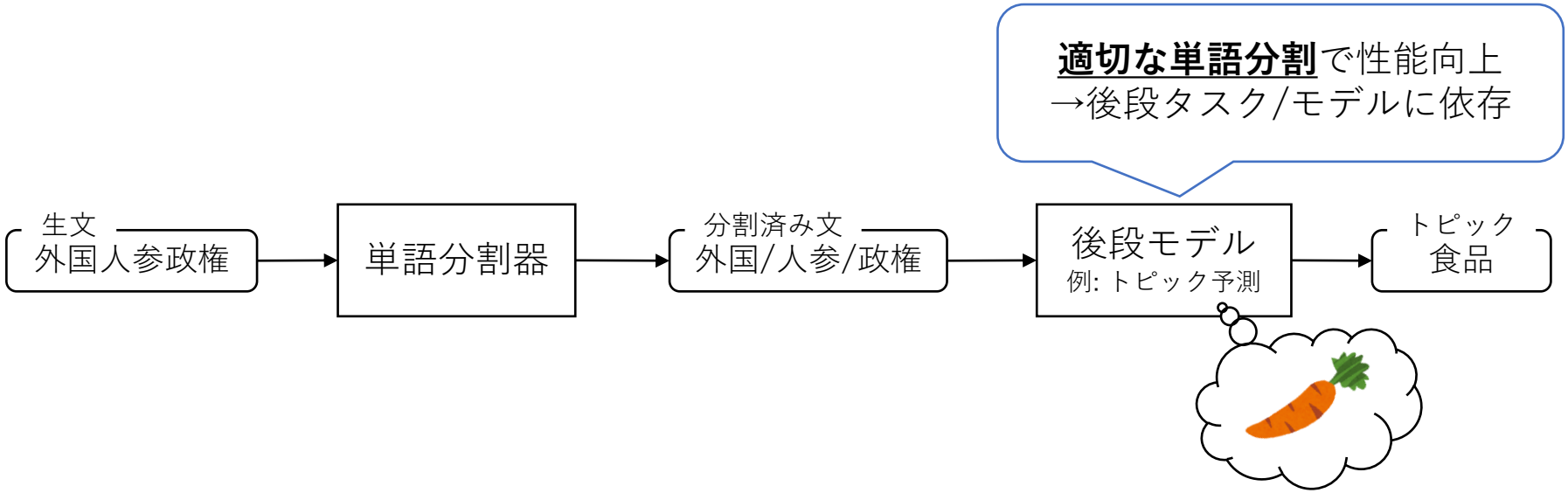
どんな話？

単語分割を**タスク**に応じて**最適化**する話



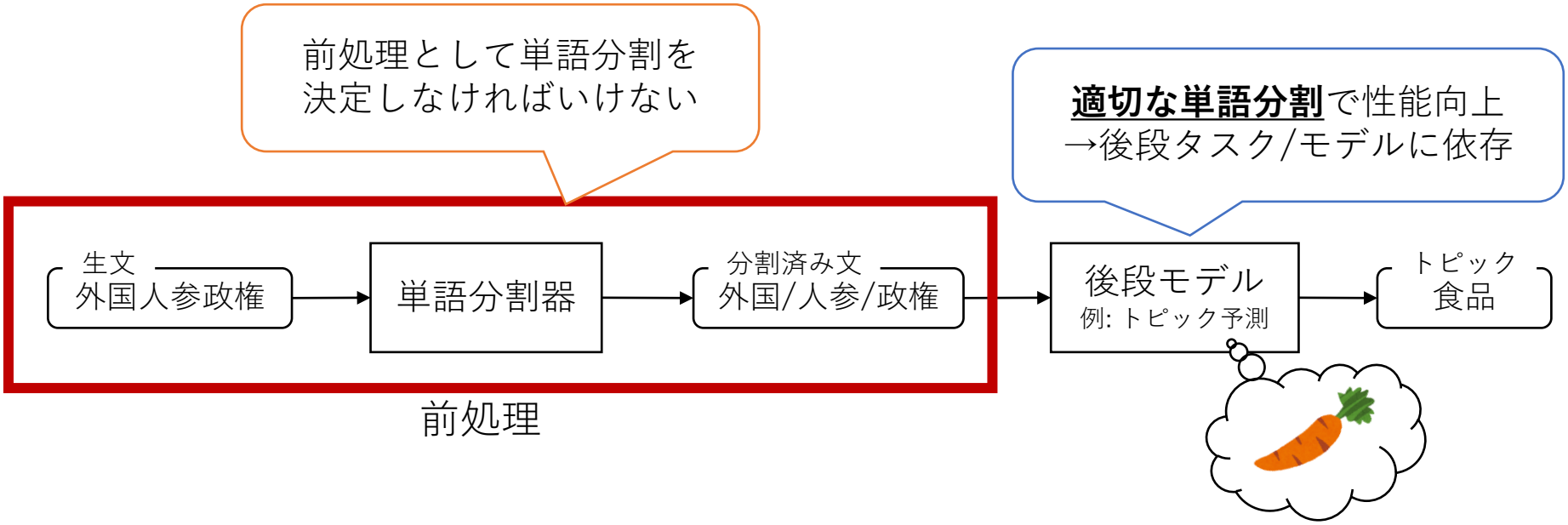
どんな話？

単語分割を**タスク**に応じて**最適化**する話



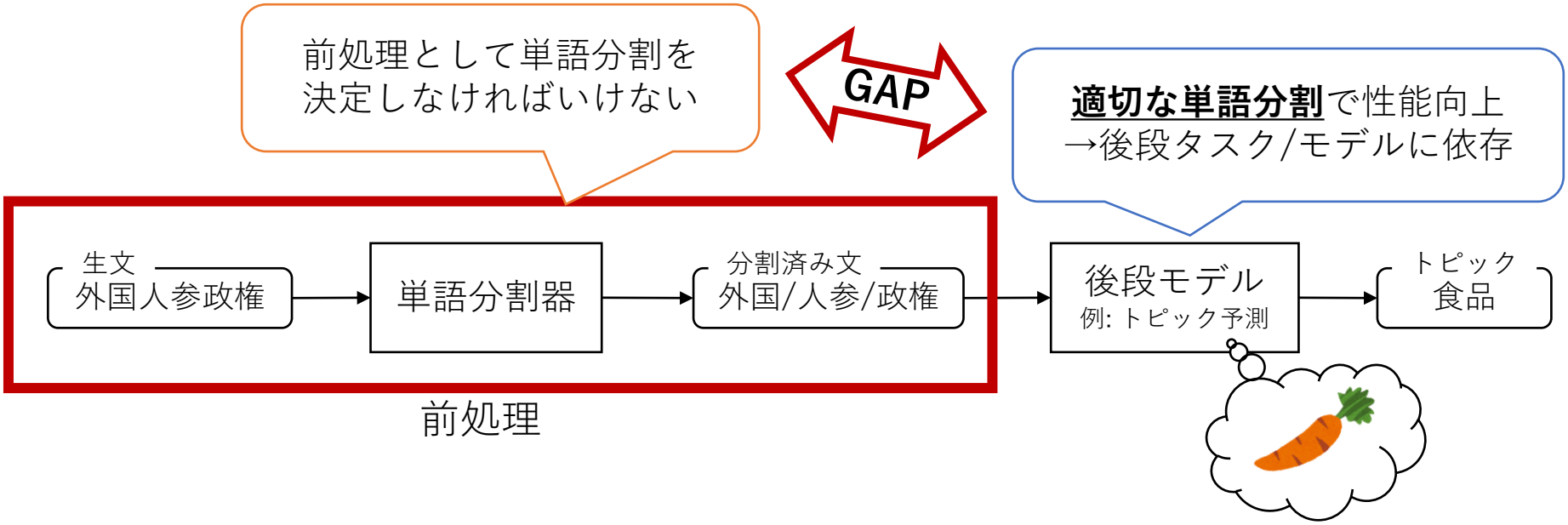
どんな話？

単語分割を**タスク**に応じて**最適化**する話



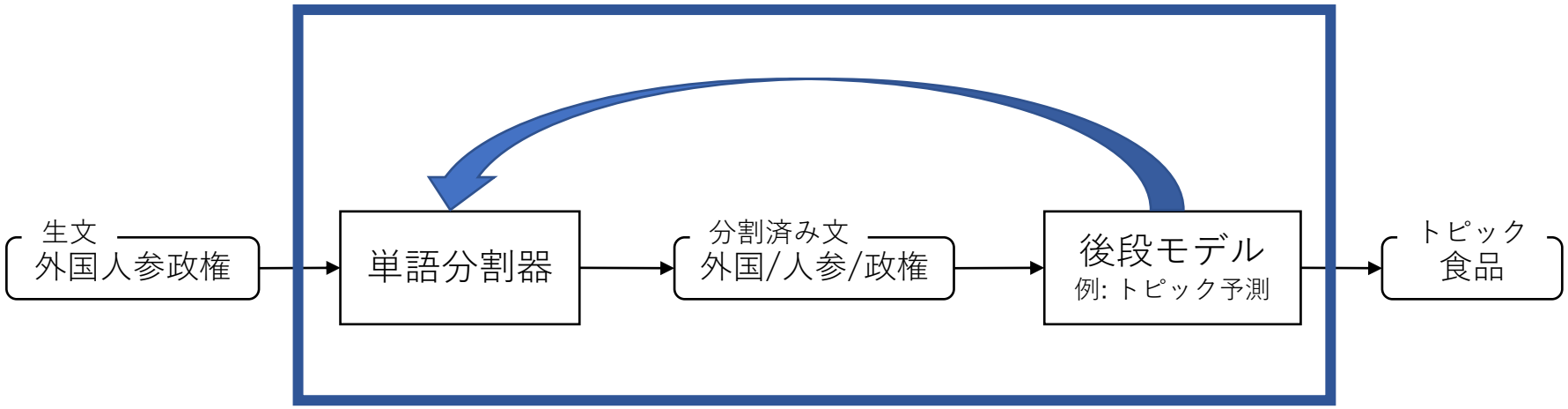
どんな話？

単語分割を**タスク**に応じて**最適化**する話



コアアイディア

単語分割を**タスク**に応じて**最適化**する話



後段モデルと同時に単語分割器を学習

嬉しさ

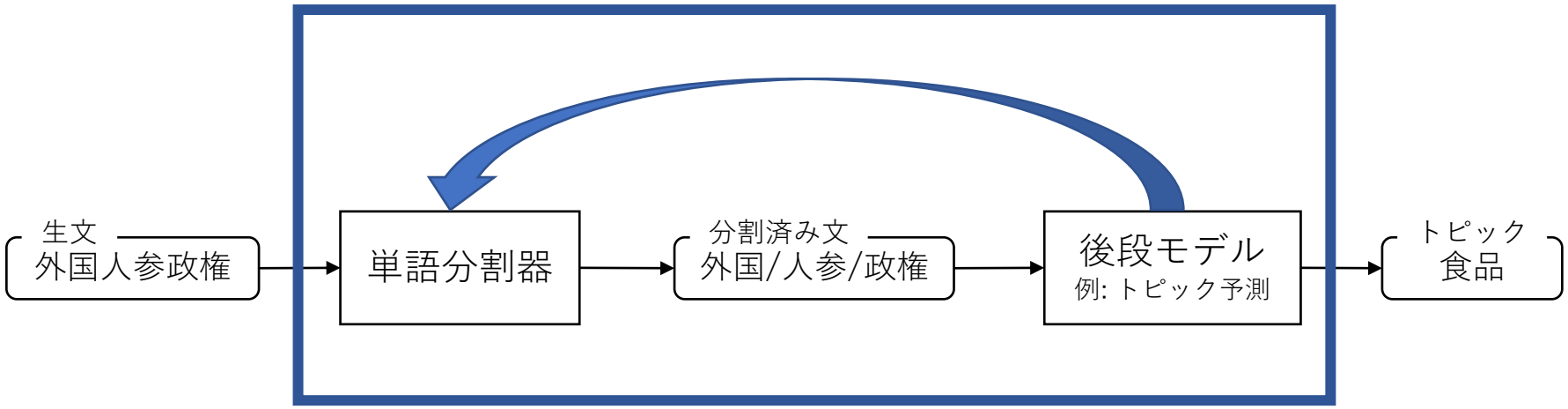
単語分割を**タスク**に応じて**最適化**する話

① タスクの**性能が上がる**と嬉しい

② タスクに**適切な単語分割が得られる**と嬉しい

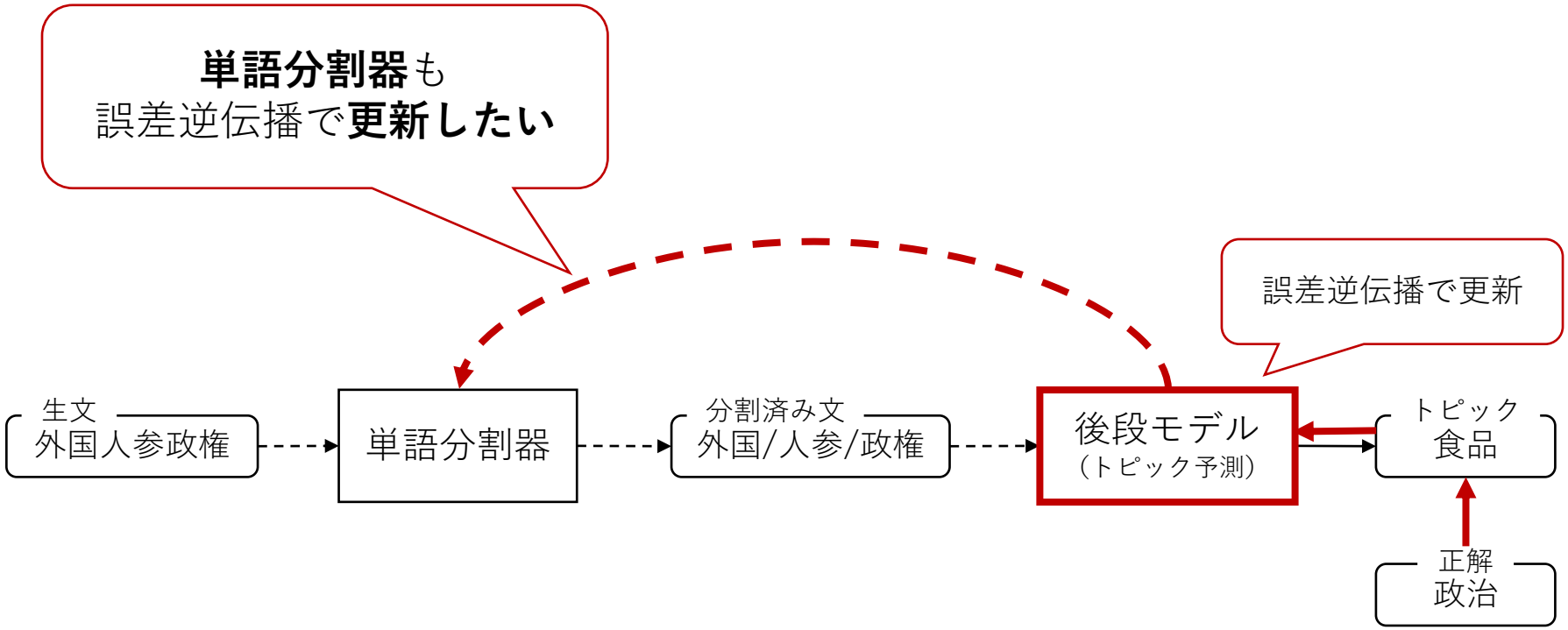
コアアイディア

単語分割を**タスク**に応じて**最適化**する話

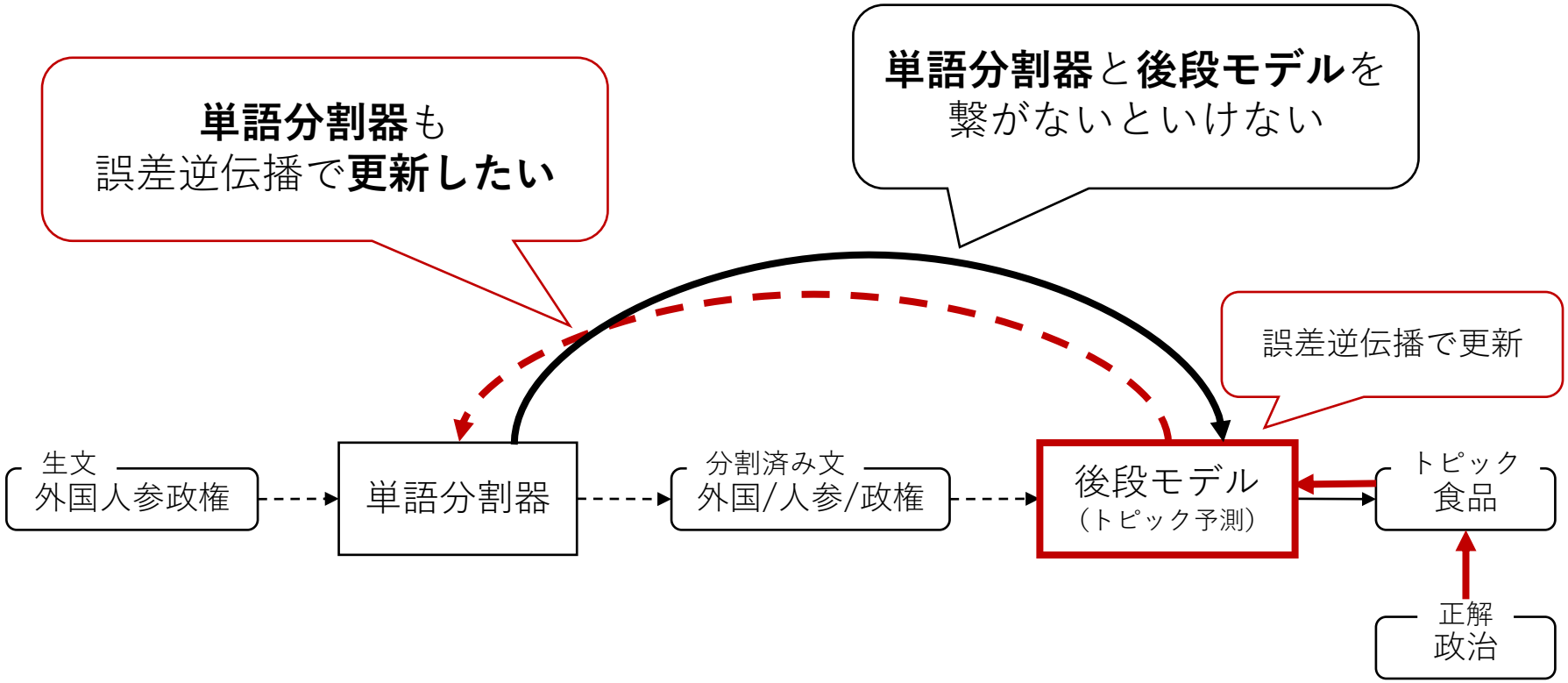


後段モデルと同時に単語分割器を学習

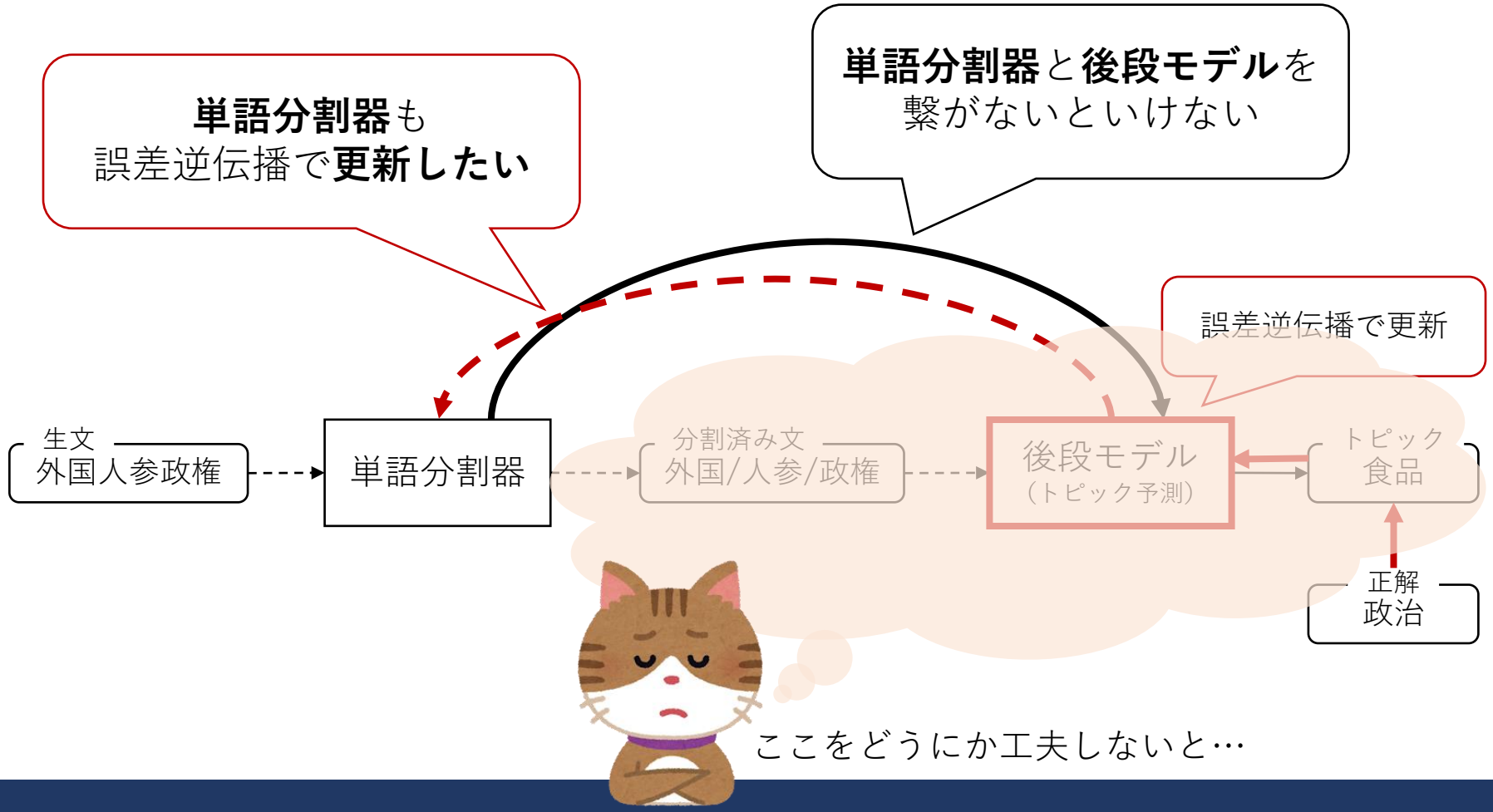
後段の学習に単語分割器を組み込みたい



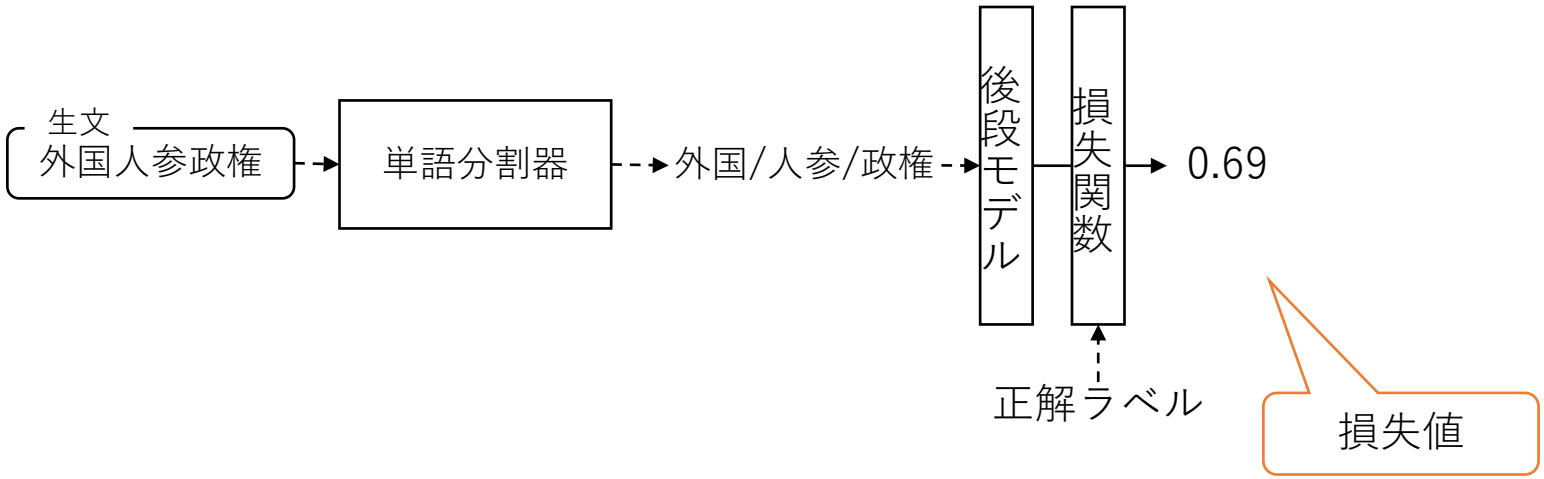
後段の学習に単語分割器を組み込みたい



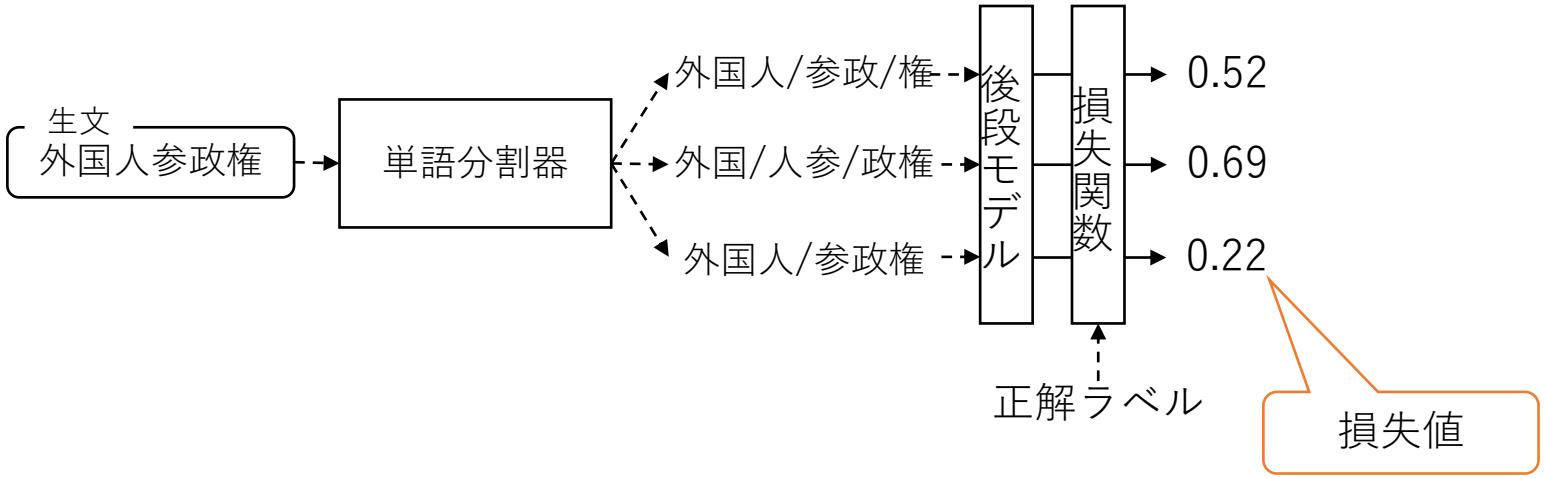
後段の学習に単語分割器を組み込みたい



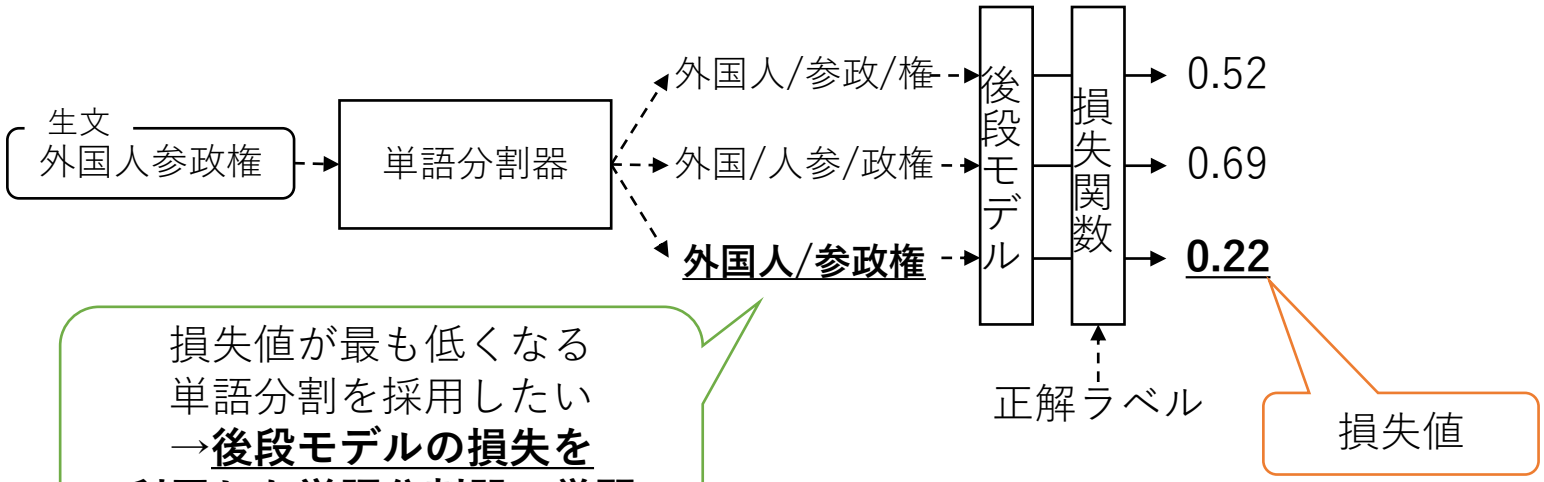
後段の学習に単語分割器を組み込む



後段の学習に単語分割器を組み込む



後段の学習に単語分割器を組み込む

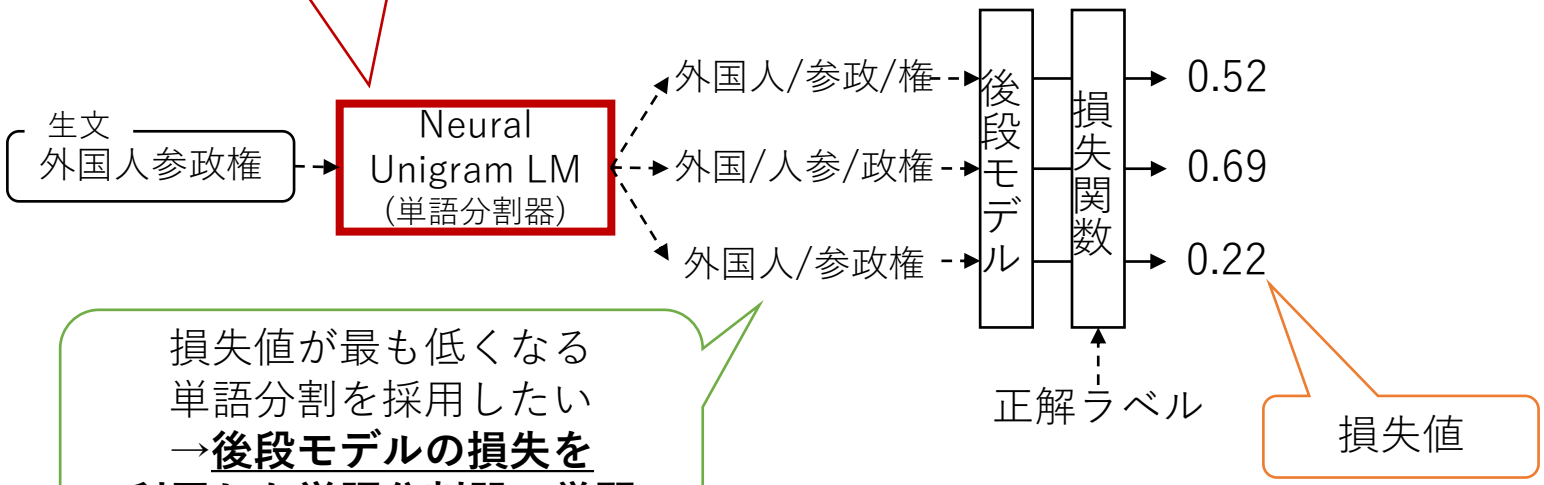


後段の学習に単語分割器を組み込む

単語分割の確率を単語確率の積で計算

$p(\text{外国}/\text{人参}/\text{政権}) = p(\text{外国})p(\text{人参})p(\text{政権})$

単語確率が学習可能パラメータ



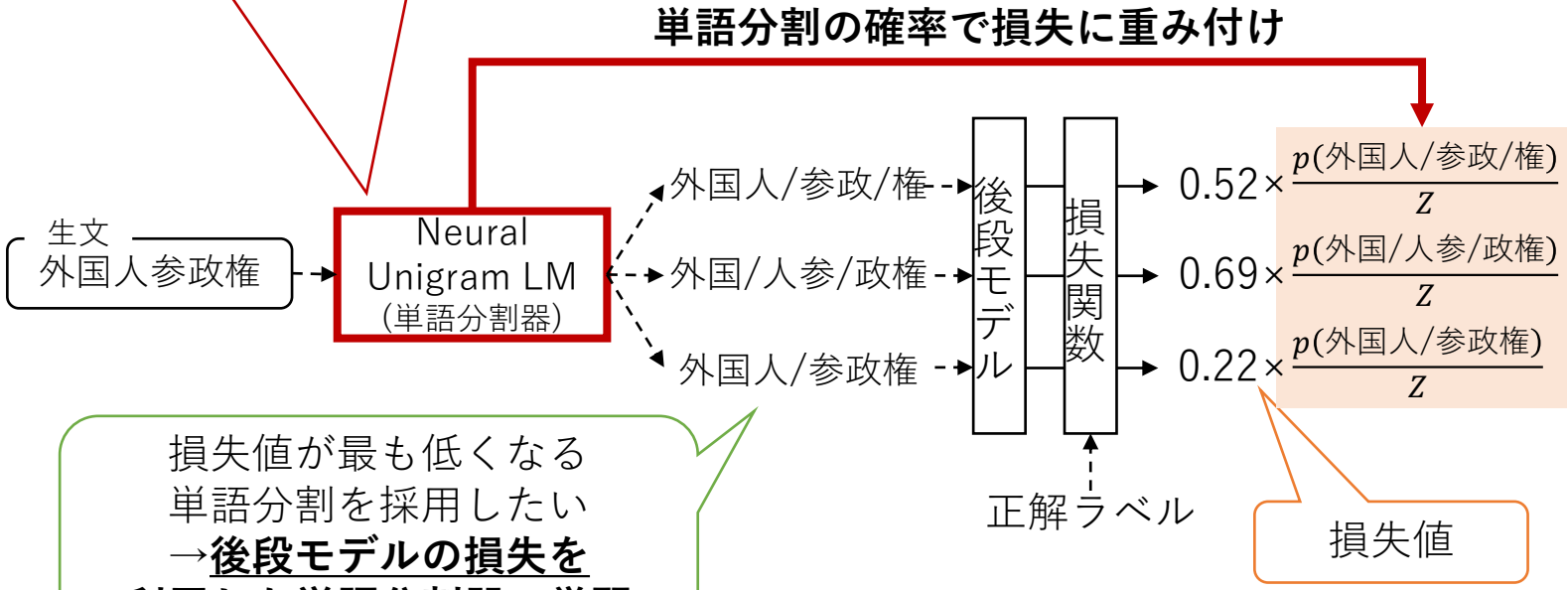
損失値が最も低くなる
単語分割を採用したい
→**後段モデルの損失を利用した単語分割器の学習**

後段の学習に単語分割器を組み込む

単語分割の確率を単語確率の積で計算

$p(\text{外国/人参/政権}) = p(\text{外国})p(\text{人参})p(\text{政権})$

単語確率が学習可能パラメータ



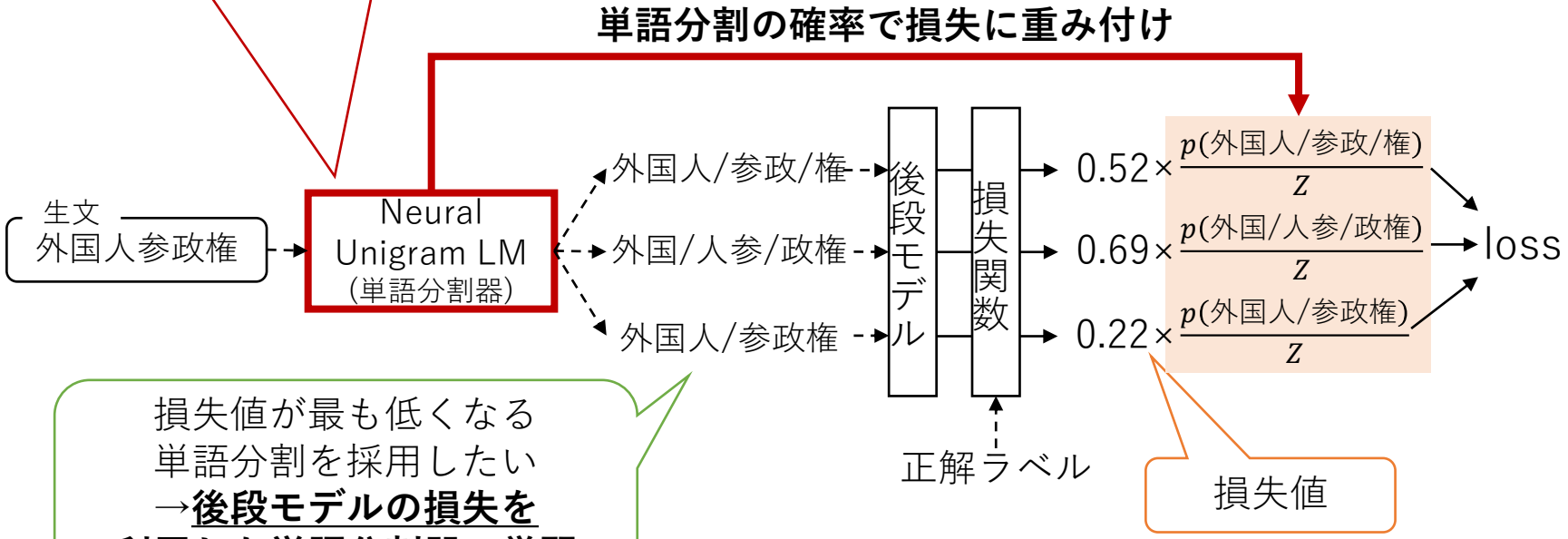
損失値が最も低くなる
単語分割を採用したい
→**後段モデルの損失を利用した単語分割器の学習**

後段の学習に単語分割器を組み込む

単語分割の確率を単語確率の積で計算

$p(\text{外国/人参/政権}) = p(\text{外国})p(\text{人参})p(\text{政権})$

単語確率が学習可能パラメータ



損失値が最も低くなる
単語分割を採用したい
→**後段モデルの損失を利用した単語分割器の学習**

後段の学習に単語分割器を組み込む

単語分割の確率を単語確率の積で計算

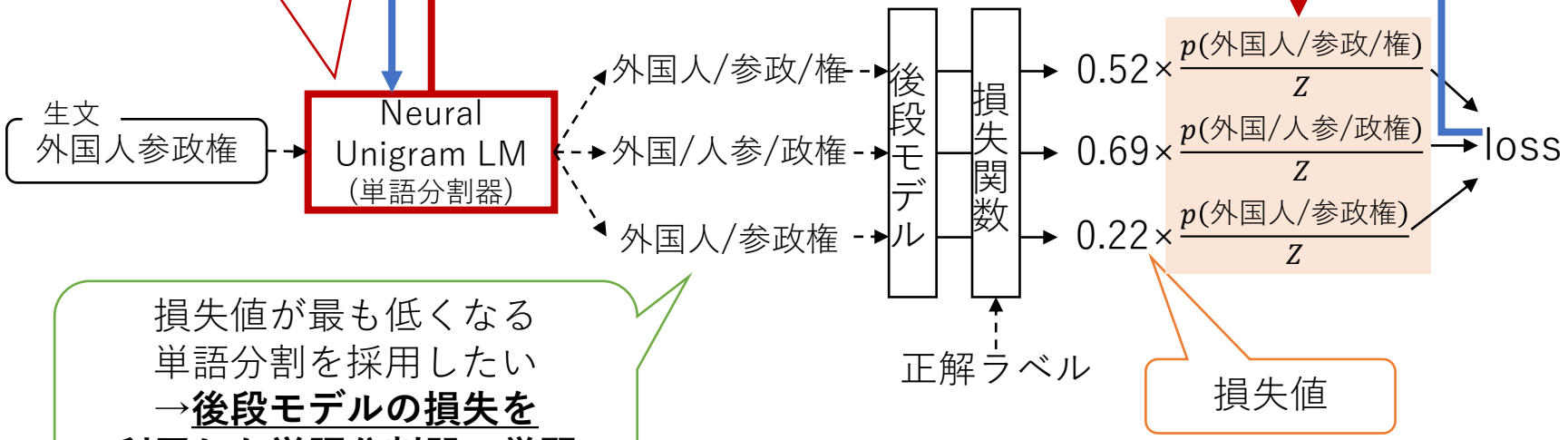
$p(\text{外国/人参/政権}) = p(\text{外国})p(\text{人参})p(\text{政権})$

単語確率が学習可能パラメータ

この損失への誤差逆伝播で
LMと後段モデルを同時に更新
→損失が小さい単語分割の
確率が上昇するように更新

勾配が計算できる

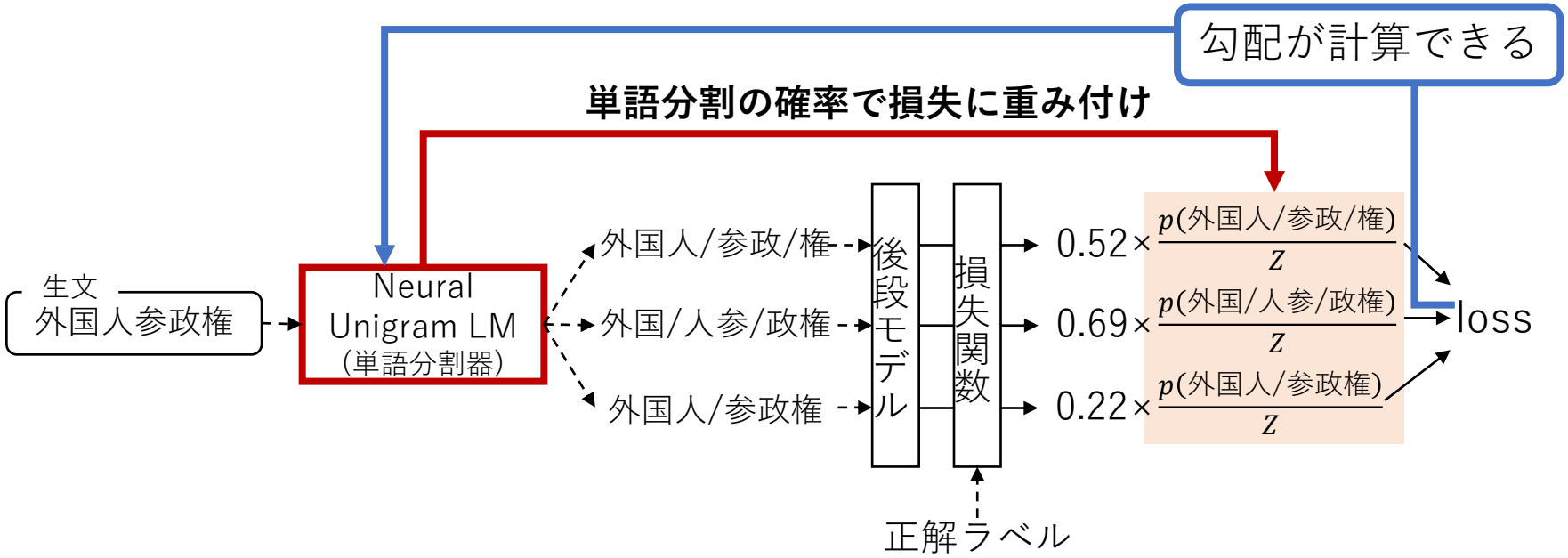
単語分割の確率で損失に重み付け



損失値が最も低くなる
単語分割を採用したい
→**後段モデルの損失を
利用した単語分割器の学習**

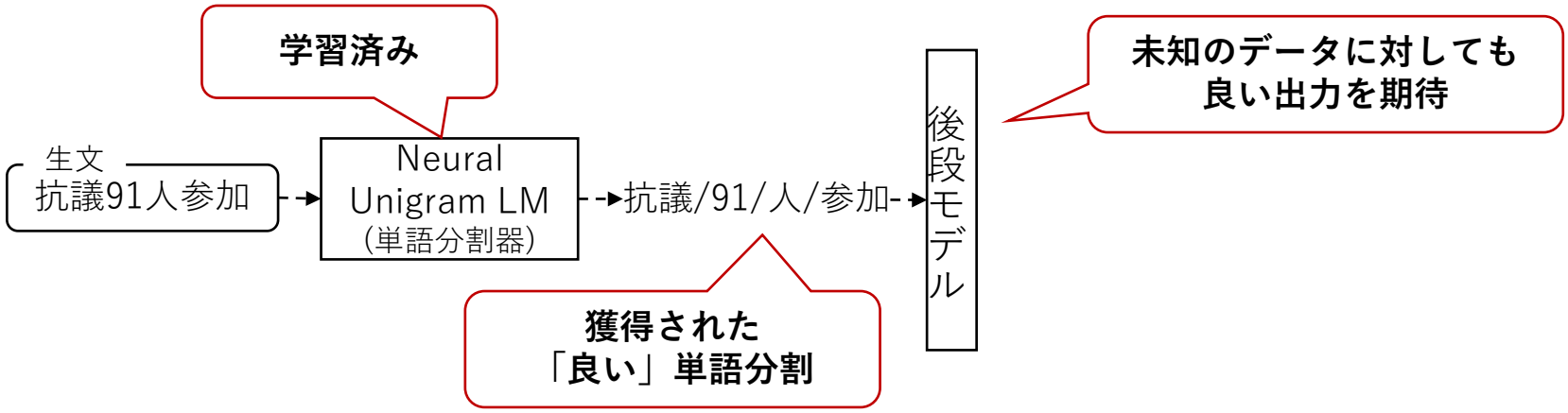
アーキテクチャの嬉しさ

- 後段モデルや損失関数は何でも良いので
  後段タスクやモデルの構造を問わず単語分割の最適化が可能
- **エンコーダー・デコーダー双方に適用すればMTでも使える**



推論時は1-best分割

学習した単語分割器は
通常のUnigramモデルと同じ速度で推論可能



実験

- 目的
  - 提案手法による単語分割器の学習が性能向上に繋がることを確かめる
  - 提案手法が複数のタスク・複数の言語で有効であることを確かめる
- 後段タスク
  - 文書分類（中国語・日本語・英語）
  - 機械翻訳（複数言語と英語のペア）
- 比較方法（後段タスクでの性能を比較）
  - Unigramによる単語分割を用いた手法
  - Unigramによる単語分割をもとに最適化を行う提案手法

実験（文書分類：BiLSTM）

- 初期状態の単語分割に比べて性能が向上（F1，5回試行平均）

タスク	言語	Unigram	Ours
感情分析	中	92.79	**93.06**
	日	86.51	**86.92**
	英	77.31	**78.88**
レビューのジャンル予測	中	47.95	**48.41**
	日	49.84	**50.79**
	英	71.68	**71.83**
レビューのレート予測	中	49.41	**49.76**
	日	53.43	**53.69**
	英	67.53	**67.90**
SNLI	英	76.75	**77.05**

実験（文書分類：BiLSTM）

- 初期状態の単語分割に比べて性能が向上（F1，5回試行平均）

タスク	言語	Unigram	Ours
感情分析	中	92.79	**93.06**
Twitter / Weibo	日	86.51	**86.92**
	英	77.31	**78.88**
レビューのジャンル予測	中	47.95	**48.41**
	日	49.84	**50.79**
	英	71.68	**71.83**
レビューのレート予測	中	49.41	**49.76**
	日	53.43	**53.69**
	英	67.53	**67.90**
SNLI	英	76.75	**77.05**

実験（文書分類：BiLSTM）

- 初期状態の単語分割に比べて性能が向上（F1，5回試行平均）

タスク	言語	Unigram	Ours
感情分析	中	92.79	**93.06**
Twitter / Weibo	日	86.51	**86.92**
	英	77.31	**78.88**
レビューのジャンル予測	中	47.95	**48.41**
JD-com, Rakuten, Amazon	日	49.84	**50.79**
	英	71.68	**71.83**
レビューのレート予測	中	49.41	**49.76**
JD-com, Rakuten, Amazon	日	53.43	**53.69**
	英	67.53	**67.90**
SNLI	英	76.75	**77.05**

実験（文書分類：BiLSTM）

- 初期状態の単語分割に比べて性能が向上（F1，5回試行平均）

タスク	言語	Unigram	Ours
感情分析	中	92.79	**93.06**
Twitter / Weibo	日	86.51	**86.92**
	英	77.31	**78.88**
レビューのジャンル予測	中	47.95	**48.41**
JD-com, Rakuten, Amazon	日	49.84	**50.79**
	英	71.68	**71.83**
レビューのレート予測	中	49.41	**49.76**
JD-com, Rakuten, Amazon	日	53.43	**53.69**
	英	67.53	**67.90**
SNLI 入力が2文	英	76.75	**77.05**

実験（機械翻訳: Transformer）

- 提案手法による性能向上を確認（BLEU）

データセット	言語対	Unigram Unigram	Ours Unigram	Unigram Ours	Ours Ours
IWSLT15	Vi→En	28.78	29.34	**29.69**	29.44
	En→Vi	31.60	31.41	**31.74**	31.70
	Zh→En	21.17	21.63	21.65	**21.89**
	En→Zh	15.25	15.45	**15.59**	15.31
WMT14	De→En	31.89	**32.19**	31.98	31.90
	En→De	27.41	**27.62**	27.52	27.44

実験（機械翻訳: Transformer）

- 提案手法による性能向上を確認（BLEU）

ソース側の単語分割
ターゲット側の単語分割

データセット	言語対	Unigram	Ours	Unigram	Ours
		Unigram	Unigram	Ours	Ours
IWSLT15	Vi→En	28.78	29.34	**29.69**	29.44
	En→Vi	31.60	31.41	**31.74**	31.70
	Zh→En	21.17	21.63	21.65	**21.89**
	En→Zh	15.25	15.45	**15.59**	15.31
WMT14	De→En	31.89	**32.19**	31.98	31.90
	En→De	27.41	**27.62**	27.52	27.44

実験（機械翻訳: Transformer）

- 提案手法に。EU）

例：ソース側の単語分割のみを提案手法で最適化

ソース側の単語分割
ターゲット側の単語分割

データセット	言語対	Unigram	Ours	Unigram	Ours
		Unigram	Unigram	Ours	Ours
IWSLT15	Vi→En	28.78	29.34	**29.69**	29.44
	En→Vi	31.60	31.41	**31.74**	31.70
	Zh→En	21.17	21.63	21.65	**21.89**
	En→Zh	15.25	15.45	**15.59**	15.31
WMT14	De→En	31.89	**32.19**	31.98	31.90
	En→De	27.41	**27.62**	27.52	27.44

実験（機械翻訳: Transformer）

- 提案手法に。 EU）

例：ソース側の単語分割のみを提案手法で最適化

ソース側の単語分割	
ターゲット側の単語分割	

データセット	言語対	Unigram / Unigram	Ours / Unigram	Unigram / Ours	Ours / Ours
IWSLT15	Vi→En	28.78	29.34	**29.69**	29.44
	En→Vi	31.60	31.41	**31.74**	31.70
	Zh→En	21.17	21.63	21.65	**21.89**
	En→Zh	15.25	15.45	**15.59**	15.31
WMT14	De→En	31.89	**32.19**	31.98	31.90
	En→De	27.41	**27.62**	27.52	27.44

ベースライン

※ベースラインを超える数値

実験（機械翻訳: Transformer）

- 提案手法に。 EU）

例：ソース側の単語分割
のみを提案手法で最適化

ソース側の単語分割
ターゲット側の単語分割

データセット	言語対	Unigram Unigram	Ours Unigram	Unigram Ours	Ours Ours
IWSLT15	Vi→En	28.78	29.34	**29.69**	29.44
	En→Vi	31.60	31.41	**31.74**	31.70
	Zh→En	21.17	21.63	21.65	**21.89**
	En→Zh	15.25	15.45	**15.59**	15.31
WMT14	De→En	31.89	**32.19**	31.98	31.90
	En→De	27.41	**27.62**	27.52	27.44

ベースライン

※ベースラインを超える数値

ターゲット側に提案手法を
用いると性能が高い傾向がある

実験（機械翻訳: Transformer）

- 提案手法に。EU）

例：ソース側の単語分割
のみを提案手法で最適化

	ソース側の単語分割
	ターゲット側の単語分割

データセット	言語対	Unigram Unigram	Ours Unigram	Unigram Ours	Ours Ours
IWSLT15	Vi→En	28.78	29.34	**29.69**	29.44
	En→Vi	31.60	31.41	**31.74**	31.70
	Zh→En	21.17	21.63	21.65	**21.89**
	En→Zh	15.25	15.45	**15.59**	15.31
WMT14	De→En	31.89	**32.19**	31.98	31.90
	En→De	27.41	**27.62**	27.52	27.44

※ベースラインを超える数値

ターゲット側に提案手法を
用いると性能が高い傾向がある

両側に提案手法を用いると
性能は低め
→学習が安定しないためか

獲得された単語分割の比較（機械翻訳）

- ソース側の分割
  - 提案手法は接尾辞などを細かく分割する傾向

最適化なし	Student s **don** ' t **have** long **hours** of learning .
最適化あり	Student s **do n** ' t **hav e** long **hour s** of learning .
ターゲット文	学生 在 校 学习 时间 不 长 。

- ターゲット側の分割
  - 主要な接尾辞（-edなど）の分割を変更する程度

ソース文	引力 与 其它 力 分 隔 开来
最適化なし	Gra vity **separate d** away from the other force s .
最適化あり	Gra vity **separat ed** away from the other force s .

獲得された単語分割の比較（機械翻訳）

- ソース側の分割
  - 提案手法は接尾辞などを細かく分割する傾向

提案手法の系列長は
“最適化なし”の1.35倍

最適化なし	Student s **don** ' t **have** long **hours** of learning .
最適化あり	Student s **do n** ' t **hav e** long **hour s** of learning .
ターゲット文	学生 在 校 学习 时间 不 长 。

- ターゲット側の分割
  - 主要な接尾辞（-edなど）の分割を変更する程度

ソース文	引力 与 其它 力 分 隔 开来
最適化なし	Gra vity **separate d** away from the other force s .
最適化あり	Gra vity **separat ed** away from the other force s .

獲得された単語分割の比較（機械翻訳）

- ソース側の分割
  - 提案手法は接尾辞などを細かく分割する傾向

提案手法の系列長は
“最適化なし”の1.35倍

最適化なし	Student s **don** ' t **have** long **hours** of learning .
最適化あり	Student s **do n** ' t **hav e** long **hour s** of learning .
ターゲット文	学生 在 校 学习 时间 不 长 。

- ターゲット側の分割
  - 主要な接尾辞（-edなど）の分割を変更する程度

提案手法の系列長は
“最適化なし”の0.99倍

ソース文	引力 与 其它 力 分 隔 开来
最適化なし	Gra vity **separate d** away from the other force s .
最適化あり	Gra vity **separat ed** away from the other force s .

獲得された単語分割の比較（機械翻訳）

- ソース側の分割
  - 提案手法は接尾辞などを細かく分割する傾向

提案手法の系列長は
“最適化なし”の1.35倍

最適化なし	Student s **don** ' t **have** long **hours** of learning .
最適化あり	Student s **do n** ' t **hav e** long **hour s** of learning .
ターゲット文	学生 在 校 学习 时间 不 长 。

- ターゲット側の分割
  - 主要な接尾辞（-edなど）の分割を変更する程度

提案手法の系列長は
“最適化なし”の0.99倍

ソース文	引力 与 其它 力 分 隔 开来
最適化なし	Gra vity **separate d** away from the other force s .
最適化あり	Gra vity **separat ed** away from the other force s .

系列長が長くなるとデコードで不利になるためか

言語・モジュールごとに単語分割の細かさが異なる

			トークン数が何倍になったか（＝分割が何倍細かくなったか）	
			ソース側だけ最適化	ターゲット側だけ最適化
ドイツ語	→	英語	2.5353	0.9992
英語	→	ドイツ語	1.3809	0.9996
ベトナム語	→	英語	1.5320	0.9993
英語	→	ベトナム語	1.4650	0.9999
中国語	→	英語	1.5175	0.9994
英語	→	中国語	1.3516	1.4713
			（ソース側の細かさ）	（ターゲット側の細かさ）

- ソース側は細かく，ターゲット側は粗く学習
  - 細かい系列（多くの短いトークンを含む）を出力するのは難しいため
- 中国語はターゲット側も細かくなっている
  - ソース側と系列の細かさを揃えるためか

言語・モジュールごとに単語分割の細かさが異なる

			トークン数が何倍になったか（＝分割が何倍細かくなったか）	
			ソース側だけ最適化	ターゲット側だけ最適化
ドイツ語	→	英語	2.5353	0.9992
英語	→	ドイツ語	1.3809	0.9996
ベトナム語	→	英語	1.5320	0.9993
英語	→	ベトナム語	1.4650	0.9999
中国語	→	英語	1.5175	0.9994
英語	→	中国語	1.3516	1.4713
			（ソース側の細かさ）	（ターゲット側の細かさ）

- **ソース側は細かく，ターゲット側は粗く学習**
  - 細かい系列（多くの短いトークンを含む）を出力するのは難しいため
- 中国語はターゲット側も細かくなっている
  - ソース側と系列の細かさを揃えるためか

言語・モジュールごとに単語分割の細かさが異なる

			トークン数が何倍になったか（＝分割が何倍細かくなったか）	
			ソース側だけ最適化	ターゲット側だけ最適化
ドイツ語	→	英語	2.5353	0.9992
英語	→	ドイツ語	1.3809	0.9996
ベトナム語	→	英語	1.5320	0.9993
英語	→	ベトナム語	1.4650	0.9999
中国語	→	英語	1.5175	0.9994
英語	→	中国語	1.3516	1.4713
			（ソース側の細かさ）	（ターゲット側の細かさ）

- **ソース側は細かく，ターゲット側は粗く学習**
  - 細かい系列（多くの短いトークンを含む）を出力するのは難しいため
- 中国語はターゲット側も細かくなっている
  - ソース側と系列の細かさを揃えるためか

言語・モジュールごとに単語分割の細かさが異なる

			トークン数が何倍になったか（＝分割が何倍細かくなったか）	
			ソース側だけ最適化	ターゲット側だけ最適化
ドイツ語	→	英語	2.5353	0.9992
英語	→	ドイツ語	1.3809	0.9996
ベトナム語	→	英語	1.5320	0.9993
英語	→	ベトナム語	1.4650	0.9999
中国語	→	英語	1.5175	0.9994
英語	→	中国語	1.3516	1.4713
			（ソース側の細かさ）	（ターゲット側の細かさ）

- ソース側は細かく，ターゲット側は粗く学習
  - 細かい系列（多くの短いトークンを含む）を出力するのは難しいため
- **中国語はターゲット側も細かくなっている**
  - ソース側と系列の細かさを揃えるためか

まとめ

- 目的：
  - 後段タスクと後段モデルに応じて適切な単語分割を探索し，
    性能向上を目指す
- 解決方策：
  - 単語分割と後段モデルを同時に最適化することで，
    後段タスクと後段モデルに応じた適切な単語分割を学習
- 貢献：
  - 提案手法はさまざまな後段タスクや後段モデルに適用可能
  - 複数言語での文書分類と機械翻訳で性能の向上に寄与
- 課題は山積：
  - 性能向上の幅が小さい
  - 学習時の計算効率が悪い

その後の発展（言語処理学会第29回年次大会 NLP2023）

- 転移学習における強化学習を用いた効率的なトークナイザとモデルの同時学習
  - 平子潤 (名大), 柴田知秀 (ヤフー)
- ニューラル機械翻訳における単語分割器のドメイン適応
  - 榎本大晟, 平澤寅庄, 金輝燦, 岡照晃, 小町守 (都立大)
- 語彙制約付きニューラル単語分割器を用いた後処理としての単語分割の後段タスクへの最適化
  - 平岡達也, 岩倉友哉 (富士通)
- 人間と機械学習のモデルそれぞれに扱いやすいトークン分割に関する実験と考察
  - 平岡達也, 岩倉友哉 (富士通)

人間とAIの第一言語獲得方向に発展させても面白い