

ニューラル機械翻訳のためのノイズ寛容なアンカー学習

根石将人
東京大学

吉永直樹
東京大学 生産技術研究所

概要

- 機械翻訳における学習データ中のノイズ対策として、SVMで提案されたアンカー学習手法をNMTモデルに導入
- 翻訳のBLEUスコアが僅かに向上し、分析の結果、翻訳モデルがアンカーを活用していることを確認

1. 背景

- 機械翻訳の学習データ等の大規模な教師データにはノイズが混入
- ノイズデータはモデルの性能低下を引き起こす

2. 関連研究

- [Khayrallah+, 18]は対訳データ中のノイズの種類を分類し、種類毎のニューラル機械翻訳(NMT)モデルへの影響を評価
- WMT2018-2020のノイズフィルタリングのシェアタスクで翻訳モデルや言語モデルを用いる手法が複数提案される
- [Dakwale+, 19]は通常の学習と、綺麗なデータのみで学習した翻訳モデルを教師とした蒸留学習を重み付けして学習

フィルタリング手法も蒸留手法も、目的の翻訳モデルと同程度以上の別モデルを用意するコストがかかる

3. 提案手法

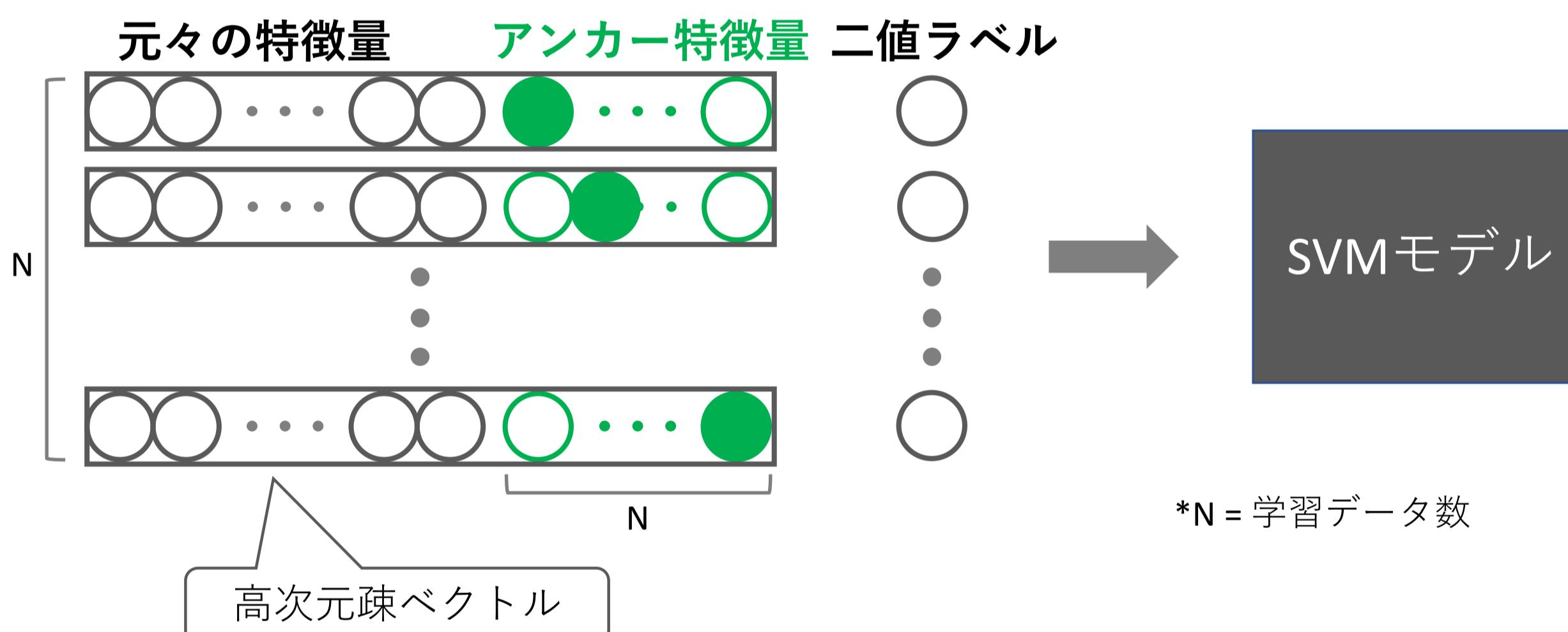
- 本研究では、ノイズデータの悪影響を減らすアンカー学習をNMTモデルの学習に導入することを提案

アンカー学習

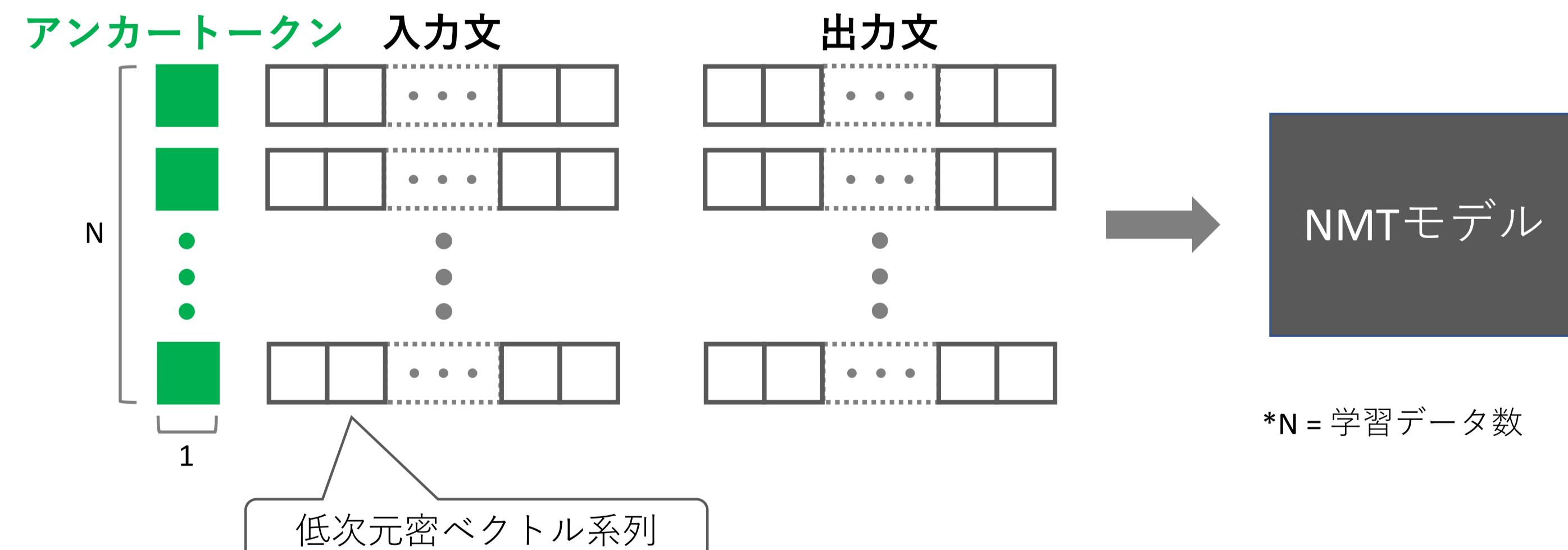
- [Goldberg+, 07]がSVMを用いた分類タスクにおいて提案
 - 各学習事例に事例固有の特徴量（アンカー）を追加
 - 通常の入力から出力の推論が難しい場合に、入力ではなくアンカーから出力を推論することが可能に
 - 誤った入出力関係の学習を避けられる

アンカー学習のNMTモデルへの適用

- SVMにおけるアンカー [Goldberg+, 07]



NMTモデルにおけるアンカー（提案手法）

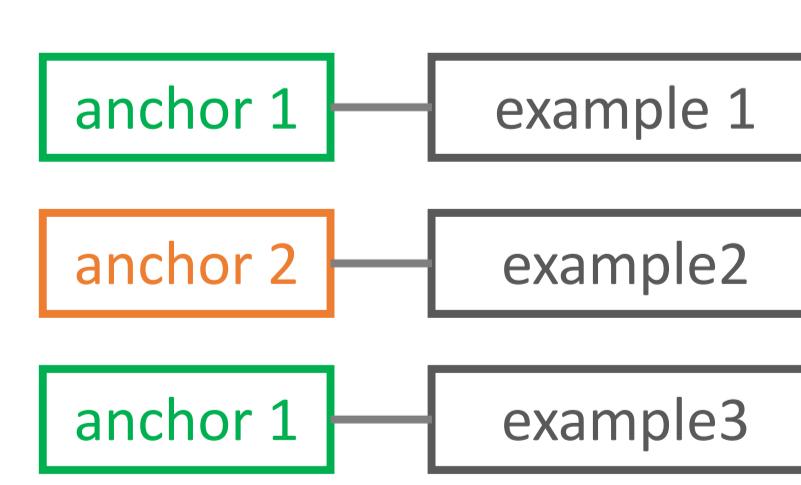


課題: 大規模な学習データではアンカートークン数が爆発

アンカーカウントを抑える工夫

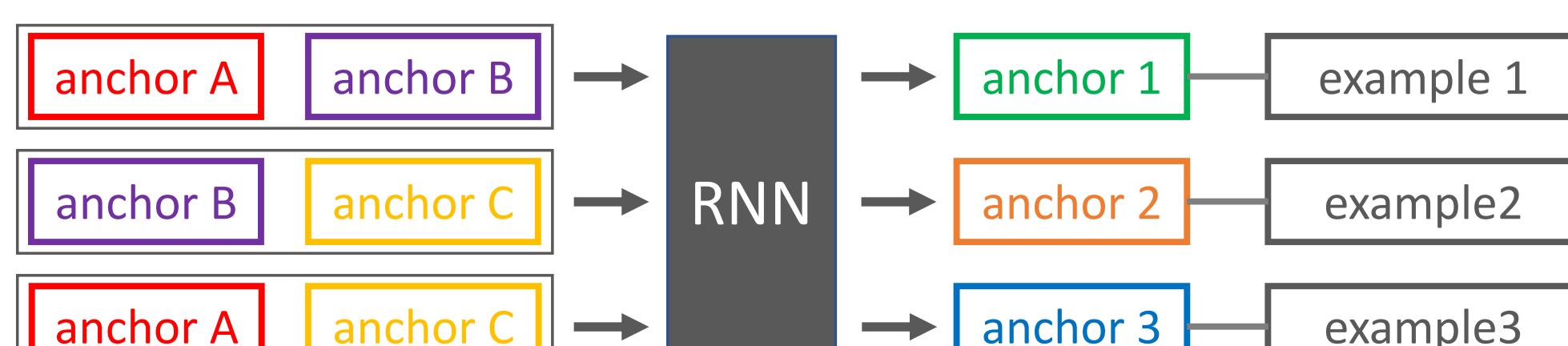
(1) ランダムグルーピング手法(RG)

- 限られた数 (N_g 個) のアンカーを用意し、複数の学習事例間での共有を許して割当
- ノイズデータが少数派であることに依拠



(2) 組み合わせ手法(Combi)

- 限られた数 (N_c 個) のアンカーを用意し、複数連ねた系列をRNNで单一のアンカーへ変換（今回、系列長は3~4）



4. 実験

- 綺麗なデータとノイズの多いデータを2:1の割合で混ぜたデータで提案手法のノイズデータへの頑健性を検証

実験設定

モデル: Transformer (相対位置自己注意機構) [Shaw+, 2018]

データ: ASPEC英日 先頭1.5M文 + 後尾750k文

WMT2017英独 5.9M文 + Paracrawl 3M文

ノイズデータ

比較手法: ランダムグルーピング手法(RG)、組み合わせ手法(Combi)
両手法共に用意するアンカーカウント (N_g , N_c) がハイパーパラメータ

実験結果: BLEUスコア (2回試行平均)

ASPEC英日	clean only	clean +noisy	# of param	WMT英独	clean only	clean +noisy	# of param
通常学習	42.19	41.00	69M	通常学習	30.05	31.11	106M
(1) RG $N_g = 10k$	42.55	42.00	74M	(1) RG $N_g = 30k$	30.27	31.43	121M
$N_g = 100k$	42.34	41.75	120M	$N_g = 300k$	30.39	31.17	259M
$N_g = 1M$	42.36	41.44	485M	$N_g = 3M$	30.26	31.08	1.53B
(2) Combi $N_c = 128$	42.27	41.90	70M	(2) Combi $N_c = 128$	30.17	31.33	107M
$N_c = 256$	42.54	41.71	70M	$N_c = 256$	30.22	31.19	107M
$N_c = 512$	42.50	41.50	71M	$N_c = 512$	30.34	31.10	108M

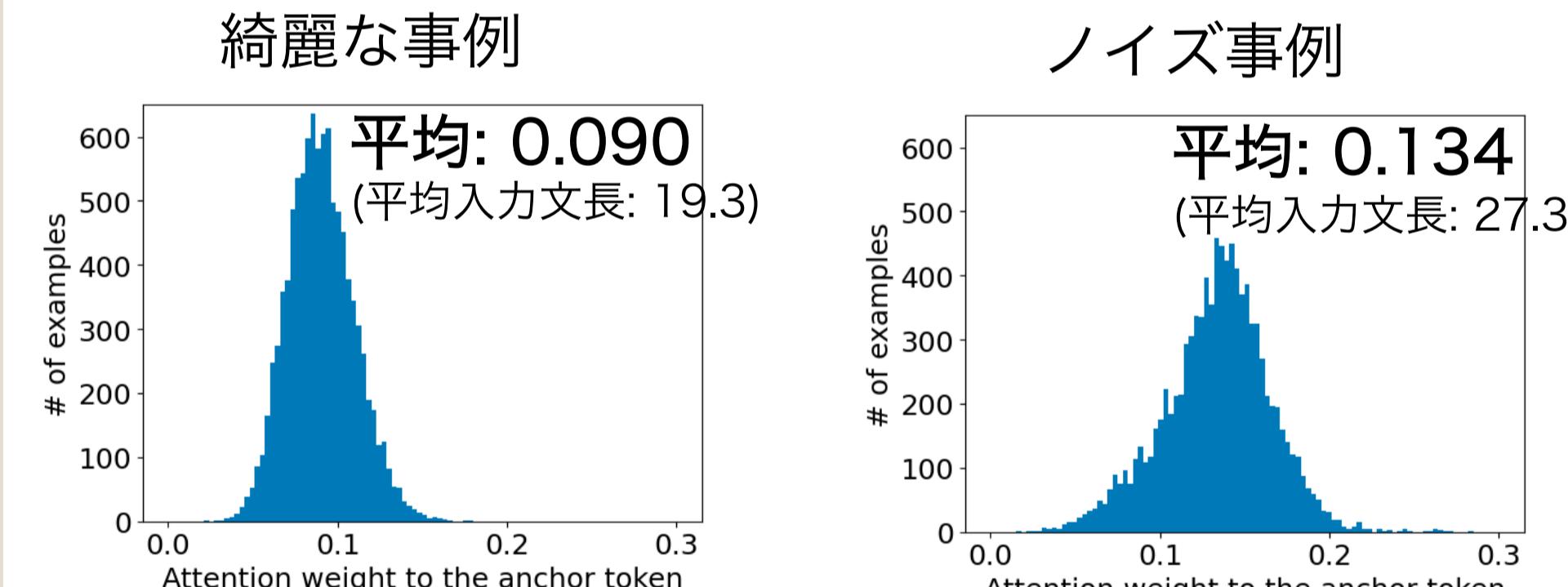
- 提案手法の両方でスコアが微増
- 組み合わせアンカーはモデルのパラメタ数増加を抑制

5. 分析

- 上の表の色付きのモデルを分析
- 学習時においてモデルがアンカーを活用しているのか確認する
- 学習データの内、綺麗な事例とノイズ事例それぞれ約1万件で、学習時のモデルの挙動を分析

分析1: Enc-Dec間アテンション(最終層)のアンカーへの重み

- アンカーが出力単語の決定にどれほど影響しているか

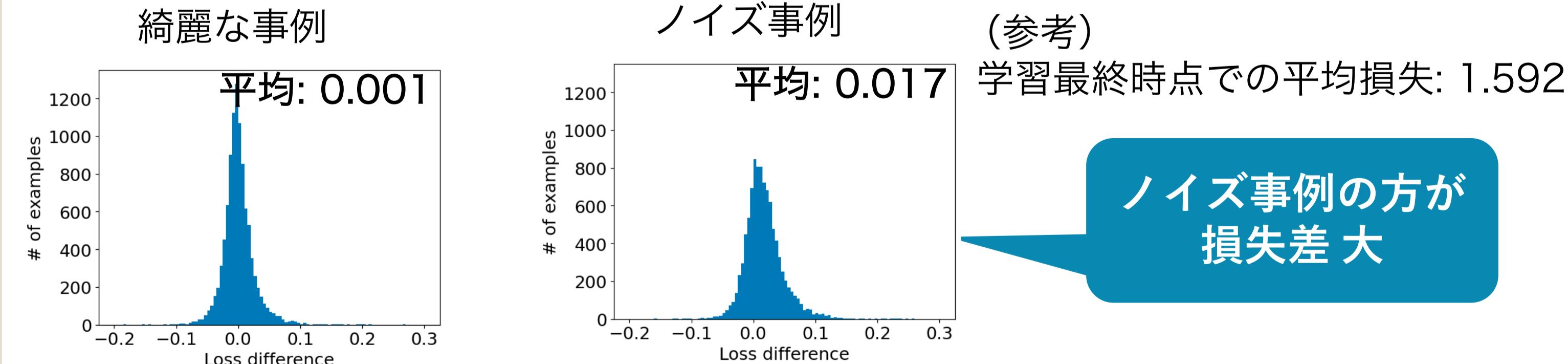


ノイズ事例の方が
アンカーへの
アテンション大

- 上位100件では入力文長が極端に短いものがほとんどで、一部 misalignment が見られた

分析2: アンカー無しと有りとの損失の差

- モデルがアンカーに頼った学習を行っているか



ノイズ事例の方が
損失差大

5. 今後の課題

- 翻訳性能の向上幅の拡大
- 提案手法のノイズの種類毎の有効性の検証

参考文献

- Khayrallah+, On the Impact of Various Types of Noise on Neural Machine Translation, NLT2018
- Dakwale+, Improving Neural Machine Translation Using Noisy Parallel Data through Distillation, MTSummit2019
- Goldberg+, SVM Model Tampering and Anchored Learning: A Case Study in Hebrew NP Chunking, ACL2007
- Shaw+, Self-Attention with Relative Position Representations, NAACL2018