音声認識と機械翻訳のランク学習による同時最適化

大串 正矢　*Graham Neubig*　*Sakriani Sakti*　**戸田 智基**　**中村 哲**

奈良先端科学技術大学院大学 情報科学研究科
〒 630-0192 奈良県生駒市高山町 8916-5
{masaya-o,neubig}@is.naist.jp

1. はじめに

音声翻訳システムは音声認識と機械翻訳，テキスト音声合成の３つのモジュールによって構成されている．たいていの場合，音声認識部は単語誤り率 (WER) が最小になるように最適化されている．しかし WER の減少が直接翻訳精度の向上に繋がる保証はない．先行研究では音声認識と機械翻訳を翻訳の指標である BLEU[1] の値が最大になるように同時最適化を行い，一定の成果を上げている [2]．また機械翻訳の最適化では，多くの素性を用いることが翻訳精度向上に繋がることが確認されているが [3]，同時最適化において多くの素性を用いる試みはまだない．そこで，多くの素性を最適化可能な対ランク最適化 PRO[3] を用いて同時最適化を行った．機械翻訳の素性と音声認識の素性に加えて，認識単語の頻度を素性として適用し，多数の素性を用いて同時最適化の効果を検証した．旅行会話 (BTEC) による音声認識，機械翻訳における実験結果より同時最適化において PRO は MERT に比べて精度が低いことが分かった．これは Brevity Penalty が大きくかかるような候補が選ばれたためである．しかし素性を多く用いることで一定の精度向上を確認することができた．

2. 音声翻訳システム

音声翻訳では，入力信号$\boldsymbol{X}$は音声認識によって原言語文の認識仮説$\boldsymbol{F}$へ変換される．さらに認識仮説$\boldsymbol{F}$ が機械翻訳へ入力され，翻訳仮説$\boldsymbol{E}$が得られる．

2.1. 音声認識の定式化

入力信号 $\boldsymbol{X}$ に対する認識仮説 $\boldsymbol{F}$ の事後確率は $P(\boldsymbol{F} \mid \boldsymbol{X})$ と定義され，下記の対数線形モデルで与えられる [4]．

$$P(\boldsymbol{F}|\boldsymbol{X}, \boldsymbol{\lambda_{ASR}}) = \frac{1}{Z_F}\exp\left\{\sum_i \lambda_i \varphi_i(\boldsymbol{F}, \boldsymbol{X})\right\} \quad (1)$$

$$Z_F = \sum_{F,X} \exp\left\{\sum_i \lambda_i \varphi_i(\boldsymbol{F}, \boldsymbol{X})\right\} \quad (2)$$

Z_F は正規化係数を表し，$\varphi_i(\boldsymbol{F}, \boldsymbol{X})$ は素性関数を表す．認識で用いられる素性は音響モデルの対数確率 $\log P_{am}(\boldsymbol{X}|\boldsymbol{F})$，原言語モデルの対数確率 $\log P_{slm}(\boldsymbol{F})$，単語数 $|\boldsymbol{F}|$ である．入力 $\boldsymbol{X}$ に対してデコーディングによって得られる確率最大の認識仮説は

$$\hat{\boldsymbol{F}} = \mathrm{argmax}_{\boldsymbol{F}}(P(\boldsymbol{F} \mid \boldsymbol{X}, \boldsymbol{\lambda_{ASR}})) \quad (3)$$

となる．

2.2. 機械翻訳の定式化

認識仮説 $\boldsymbol{F}$ に対する翻訳結果 $\boldsymbol{E}$ の事後確率は $P(\boldsymbol{E} \mid \boldsymbol{F})$ と定義され，下記の対数線形モデルで与えられる．

$$P(\boldsymbol{E} \mid \boldsymbol{F}, \boldsymbol{\lambda_{MT}}) = \frac{1}{Z_E}\exp\left\{\sum_{i'} \lambda_{i'} \varphi_{i'}(\boldsymbol{E}, \boldsymbol{F})\right\} \quad (4)$$

Z_E は式 (2) と同じく正規化係数である．$\varphi_{i'}(\boldsymbol{E}, \boldsymbol{F})$ として翻訳で標準的に用いられる素性は言語モデル対数確率 $\log P_{lm}(\boldsymbol{E})$，翻訳モデルの前向き対数確率 $\log P(\boldsymbol{E} \mid \boldsymbol{F})$，後ろ向き対数確率 $\log P(\boldsymbol{F} \mid \boldsymbol{E})$，並べ替えモデル対数確率などの計 14 素性である [5]．$\hat{\boldsymbol{E}}$ は式 (3) と同じく確率最大の翻訳仮説を表す．

3. 重み最適化

最適化とはある制約条件のもとで目的関数を最大化もしくは最小化することである．通常，音声認識のパラメータ $\boldsymbol{\lambda_{ASR}}$ を音声認識の評価尺度である単語誤り率 (WER) で最適化する．

$$\hat{\boldsymbol{\lambda}}_{\boldsymbol{ASR}} = \mathrm{argmin}_{\boldsymbol{\lambda_{ASR}}} WER(\boldsymbol{F}_{ref}, \hat{\boldsymbol{F}}) \quad (5)$$

表 1: 認識，翻訳候補の例

認識候補		翻訳候補			
i	$f(i)$	j	$e(i,j)$	$\boldsymbol{\varphi}(i,j)$	$b(i,j)$
1	“何をお飲みになりますか”	1	“what would you like to drink ?”	[2 4]	1.00
2	“何用をお飲みになりますか”	2	“what would you like for ?”	[3 8]	0.60
3	“何いをお飲みになりますか”	3	“what do you have ?”	[6 1]	0.28

機械翻訳のパラメータ $\boldsymbol{\lambda_{MT}}$ を翻訳の評価尺度である BLEU[1] で最適化する．

$$\hat{\boldsymbol{\lambda}}_{\boldsymbol{MT}} = \mathrm{argmax}_{\boldsymbol{\lambda_{MT}}} BLEU(\boldsymbol{E}_{ref}, \hat{\boldsymbol{E}}) \tag{6}$$

最適化の手法については次節以降に示す．

3.1. 誤り率最小化学習 (MERT:Minimum Error Rate Training)

MERT[6] は評価関数を直接最大化する手法である．通常，BLEU や WER などの評価関数は局所解が存在し，微分不可能であるため厳密に最大化することは困難である．そこで，MERT は与えられた素性に対して，それに付随する全てのパラメータを最適化するのではなく，一つ一つのパラメータを順に最適化することにより評価関数を最適化する手法である．また一つのパラメータの最適化手法は Och の線形探索アルゴリズム [6] である．

3.2. 対ランク最適化 (PRO:Pairwise Rank Optimization)

MERT は線形探索を行っており，学習に必要な時間が素性の数に比例して増加するため，高次元の素性空間における最適化には適していない．そこで多くの素性に対応可能な対ランク最適化 PRO[3] を適用する．PRO はランク学習を用いた重み最適化法である．学習の段階では翻訳仮説に対して，予め BLEU スコアを計算しておく．次に翻訳候補のペアを作り，BLEU スコアが高い候補から低い候補の素性の差分を取り，プラスのラベルを与える．同様に BLEU スコアが低い候補から高い候補に対しても素性の差分を取り，マイナスのラベルを与える．翻訳のペアを仮に e(i,j) と e(i,j')，素性のペアを $\boldsymbol{\varphi}(i,j)$，$\boldsymbol{\varphi}(i,j')$ とし，それらの BLEU スコアを $b(i,j)$，$b(i,j')$ とする．例えば，表 1 の第 1 と第 3 候補において得られる評価関数は $b(1,1) > b(1,3)$ となり，素性の差分とラベルは ([-4,3],+) と ([4,-3],-) となる．これを分類器の学習データとして加える．また全ての候補対を利用するのではなく，下記式の通り BLEU の差が 0.05 以上である候補対のみ利用することで，BLEU の差が有意でない候補対をデータから取り除き，過学習を防ぐ．

$$\alpha(|b(i,j) - b(i,j')|) = \begin{cases} 0 & if\ |b(i,j) - b(i,j')| < 0.05 \\ 1 & otherwise \end{cases} \tag{7}$$

3.3. デコーディングと同時最適化

通常，音声認識と機械翻訳は個別に最適化されており，認識の確率が最も高い候補を翻訳し，翻訳の確率が最も高い候補を出力する．しかし，認識の確率が最も高い候補が翻訳に適しているとは限らないため認識において複数の候補を考慮すると更なる精度向上が実現できる [7]．同時デコーディングにおける事後確率を対数線形モデルで表すと下記の用になる．$Z_{E,F}$ は式 2 と同じく正規化係数である．

$$P(\boldsymbol{E}, \boldsymbol{F}|\boldsymbol{X}) = \frac{1}{Z_{E,F}} \exp\left\{\sum_i \lambda_i \varphi_i(\boldsymbol{E}, \boldsymbol{F}, \boldsymbol{X})\right\} \tag{8}$$

入力信号 $\boldsymbol{X}$ に対してデコーディングにより得られる確率の最も高い翻訳仮説は

$$\hat{\boldsymbol{E}} = \mathrm{argmax}_{\boldsymbol{E}}(P(\boldsymbol{E}, \boldsymbol{F} \mid \boldsymbol{X}, \boldsymbol{\lambda_{ASR}}, \boldsymbol{\lambda_{MT}})) \tag{9}$$

本研究では同時デコーディングの手法として認識の n-best を用いて翻訳を行う．つまり式 (9) の確率最大の仮説を探索する時，$\boldsymbol{F}$ の母集団は認識の n-best であり，$\boldsymbol{E}$ の母集団は各認識仮説に対して生成される翻訳候補である．同時最適化では選ばれた仮説と参照文の BLEU スコアが最も高くなるようなパラメータ $\boldsymbol{\lambda_{ASR}}$，$\boldsymbol{\lambda_{MT}}$ を同時に最適化する．

$$\hat{\boldsymbol{\lambda}}_{\boldsymbol{ASR}}, \hat{\boldsymbol{\lambda}}_{\boldsymbol{MT}} = argmax_{\boldsymbol{\lambda_{ASR}}, \boldsymbol{\lambda_{MT}}}(BLEU(\boldsymbol{E}_{ref}, \hat{\boldsymbol{E}})) \tag{10}$$

パラメータの最適化手法は先行研究では上記に記した MERT であり，本研究ではこれに加えて PRO の効果を検証する．これらの手法により，通常の個別最適化では得られなかった候補が選ばれるように同時最適化が行われる．

F： 熱 に 効く 薬 が 欲しい の です が

$\varphi_{\text{熱}}(F)=1, \varphi_{\text{に}}(F)=1, \varphi_{\text{は}}(F)=0, \varphi_{\text{薬}}(F)=1, \varphi_{\text{が}}(F)=2 \ldots$

図 1: 認識単語の頻度

表 2: 本実験で使用したデータサイズ

学習データ (音響モデル)	486 時間
学習データ (言語モデル (原言語))	12k 文
学習データ (翻訳モデル)	162k 文
学習データ (言語モデル (翻訳言語))	162k 文
チューニングデータ	610 文
テストデータ	610 文

4. 新たな素性：認識単語の頻度

PRO の利点の一つは多くの素性が利用できるところにある．そこで，新たな素性として，認識候補の単語を素性として加える．これにより翻訳に寄与する単語の認識に対する重みを最適化することが期待される．図 1 に示すように音声認識用の辞書を用意し, 単語単位で一致した個数をカウントする．

5. 実験結果と考察

5.1. 実験条件

実験で利用したデータは表 2 の通りである．HMM 音響モデルを使用し学習には CSJ の男女講演データ 486 時間を使用した．原言語と目的言語の言語モデルの学習には旅行会話コーパス BTEC のデータを原言語では 12k 文，目的言語では 162k 文を使用し Trigram 言語モデルを使用した．翻訳モデルの学習には BTEC162k 文の日本語と英語のパラレルコーパスを使用し，翻訳モデルは Moses[5] によって構築されたフレーズベース翻訳モデルである．チューニングは BTEC の 610 文の日本語の音声データを使用し，テストはチューニングでは使用していない BTEC の 610 文の日本語の音声データを使用した．なお学習データ及びチューニングデータ，テストデータの文末からハテナ，コンマを削除している．理由として音声認識の際に文末に句読点が含まれず，翻訳精度に影響を及ぼすと考えられたためである．

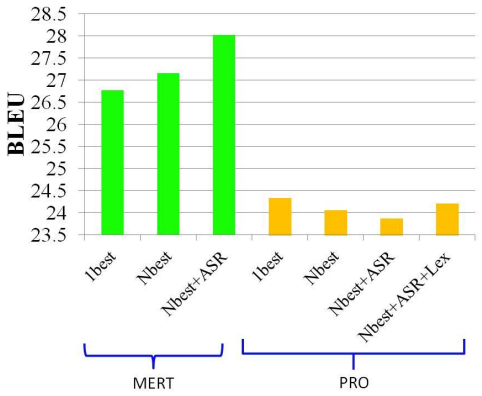

図 2: チューニングデータにおける翻訳精度

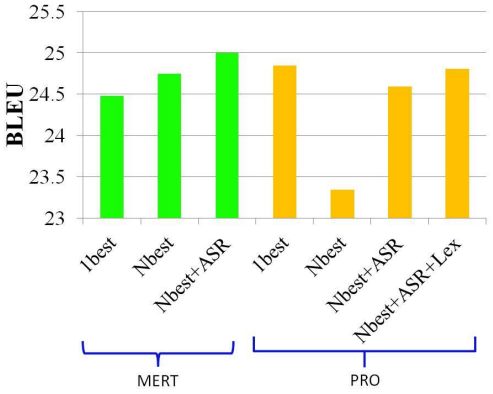

図 3: テストデータにおける翻訳精度

5.2. チューニング条件

MERT と PRO を最適化手法として使用し，音声認識で出力される認識仮説は 1-best と 50-best を使用した．翻訳の素性はベースラインで通常の 14 素性を使用した．さらに認識の n-best を考慮したシステムでは，2.1 節で説明した 3 素性を加えたシステムを「+ASR」と記し，4 節で説明した語彙素性を加えたシステムを「+Lex」と記する．MERT と PRO では Moses[5] の実装を使用し，PRO に用いる 2 値識別器として MegaM[8] を使用した．スコアは 3 回の平均値を用いて，翻訳仮説は 50 として，イテレーション回数は 25 回とした．

5.3. 実験結果と考察

実験結果を図 2，図 3 に示す．MERT において認識の n-best を考慮した方が BLEU スコアが上昇している．理由として n-best から翻訳精度向上に寄与するような候補を選択可能となったためである．PRO において認識の n-

表 3: フィラーに対して学習された重み

フィラー	重み	フィラー	重み
えっ	-0.0031	ああ	0.0001149
ええ	0.00252	あっ	0.012279
え	0.05461	っ	-0.001054

表 4: 最大と最小の学習された重み

認識単語	重み	認識単語	重み
が	0.0983	か	-0.0835
ん	0.0864	何	-0.0789
い	0.0813	は	-0.0706

best を考慮した方が BLEU スコアが減少している．これは PRO の問題点として短い翻訳文を選んでしまうことが指摘されており [9]，この問題によって短い候補文を選び Brevity Penalty が大きくかかったことが原因だと思われる．また MERT においてチューニングデータ，テストデータ共に音響的特徴量が BLEU スコアを上昇させていることが確認できる．PRO においては n-best を利用することにより BLEU スコアが低下しており，音響的特徴量を素性として加えることにより，テストデータで BLEU スコアの上昇が見られたが，1-best の PRO の値を超えるほどではない．各手法によるチューニング時間の計測も行った．MERT は 1best で 3,360 秒，n-best で 210,000 秒，音響的特徴量を考慮した n-best では 230,000 秒の時間がかかったが，PRO では 1-best で 1,700 秒，n-best で 49,000 秒，音響的特徴量を考慮した場合で 55,000 秒と MERT に比べチューニングに要する時間は大きく減少している．本研究で提案した認識単語の頻度も BLEU スコアの上昇に寄与しており，BLEU スコアがチューニングデータにおいては 0.4、テストデータにおいては 0.2 程度の翻訳精度の向上が見られ，多くの素性を用いることが BLEU スコアの向上に寄与する可能性が確認できた．新しい素性に期待した効果は下記の 2 点である．

- フィラーなどの余分な単語の除去
- 翻訳に重要な単語が含まれる候補の優先的選択

これらの効果について分析を行った．まずフィラーに関して，フィラーと最適化によって出力された重みの関係性を表 3 に示す．フィラーと思われる認識単語の重みはマイナスになっている単語もあるが，全体的にマイナスになる傾向は見られなかった．原因として，チューニングデータが旅行会話の読み上げであるため，フィラーが少なかったのことが挙げられる． 次に翻訳に重要な単語が含まれる候補を優先的に選択する場合について述べる．表 4 より，最も重みが高くなったのが「が」のような機能語である．例として重みが高い「が」と重みの低い「は」という単語に着目してみる．この単語が含まれる場合とそうでない場合において分析を行った．その分析で発見した例を表 5 に示す．

表 5: 認識単語と効果の例

認識単語	英文
一番 小さい の が いい です	i'd like the smaller
一番 小さい の は いい です	what is the smallest

この例より「が」と「は」は機能語のため，翻訳精度に影響を及ぼす．このため良い影響を与える場合は正の重みになり，悪い影響を与える場合は負の重みになる．タスクによって重要な単語が変わることが確認できた．

6. Conclusion

本研究では新たな素性を加え，PRO による同時最適化の検討を行った．実験結果より PRO はチューニングにおいて MERT に比べると精度が劣るが，多くの素性を用いることで精度が向上することが確認できた．今後は素性とデータ数を増加し PRO の効果を検証していきたい．また MIRA[10] など PRO 以外に多くの素性に対応可能な学習法が存在するため，これらの同時最適化を用いた際の効果を検証したい．

参考文献

[1] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a msthod for automatic evaluation of machine translation. In *Association for Computational Linguistics*, 2002.

[2] Xiaodong He, Li Deng, and Alex Acero. Why word error rate is not a good metric for speech recognizer training for the speech translation task. In *ICASSP*, 2011.

[3] Mark Hopkins and Jonathan May. Tuning as ranking. In *Empirical Method in Natural Language Processing*, 2011.

[4] Franz Josef Och and Hermann Ney. Discriminative training and maximum entropy models for statistical machine translation. In *Association for Computational Linguistics*, 2002.

[5] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, et al. Moses: Open source toolkit for statistical machine translation. In *Association for computational linguistics*, volume 45, page 2, 2007.

[6] Franz Josef Och. Minimum error rate training in statistical machine translation. In *Association for Computational Linguistics*, 2011.

[7] Hermann Ney. Speech translation: Coupling of recognition and translation. In *IEEE*, 1999.

[8] Hal Daumé III. Notes on CG and LM-BFGS optimization of logistic regression. `http://hal3.name/megam/`, August 2004.

[9] Preslav Nakov, Francisco Guaman, and Stephan Vogel. Optimizing for sentence-level bleu+1 yields short translation. pages 1979–1994, 2012.

[10] Taro Watanabe, Jun Suzuki, Hajime Tsukada, and Hideki Isozaki. Online large-margin training statistical machine translation. In *Empirical Methods in Natural in Natural Language Processing and Computational Natural Language Learning*, 2007.