Vol.2012-SLP-93 No.1
2012/10/26

リスク最小化学習に基づく識別的言語モデル

小林 彰夫[1,a)]　奥 貴裕[1]　藤田 悠哉[1]　佐藤 庄衛[1]　中川 聖一[2]

概要：音声認識の誤り傾向を反映した識別的言語モデルについて述べる．識別的言語モデルは，言語的な文脈により活性化する素性関数とその重みにより定められ，文仮説へのペナルティスコアとして表現される．本稿では，単語誤りに基づくリスクを尺度としたリスク最小化学習による識別的言語モデルについて，学習データに対する正解ラベルの付与の有無に応じた学習方法の観点から論じ，放送番組を対象とした音声認識実験により，提案する識別的言語モデルが統計的に頑健かつ有意な音声認識性能の改善となることを示す．

キーワード：Bayes リスク，識別学習，多目的最適化，半教師あり学習

Discriminative Language Modeling Based on Risk Minimization Training

Kobayashi Akio[1,a)]　Oku Takahiro[1]　Fujita Yuya[1]　Sato Shoei[1]　Nakagawa Seiichi[2]

Abstract: This paper describes discriminative language models (LMs) that reflect information about word errors in automatic speech recognition (ASR). The discriminative LMs are implemented as a set of penalty scores employing linguistic features and their weighting factors. The models are estimated in the basis of minimization of expected risks that are closely associated with word errors. In transcribing Japanese broadcast programs, the semi-supervised discriminative LM achieved the best results in word error rates compared with the supervised and unsupervised LMs and conventional discriminative LMs based on maximization of conditional log-likelihoods.

Keywords: Bayes risk, discriminative training, multi objective programming, semi-supervised training

1. はじめに

放送における音声認識技術は，字幕制作やメタデータ制作，放送アーカイブへの利用など，広範な放送サービスに関わっている．字幕制作支援を目的とした音声認識技術はすでに実用化されており，ニュースや情報番組を対象に運用されている [1]．また，アーカイブ等の放送コンテンツの閲覧やキーワードによる検索を目的とした報道番組の書き起こしシステムの研究も行われている [2]．音声認識技術は字幕やメタデータ制作コストの低減に貢献する一方で，放送番組におけるすべての音声に完全な文字出力を与えるものではなく，認識結果には多少の誤りが含まれる．音声認識の誤りは，雑音などの音響的な条件や話題，話者の話し方などの条件に大きく依存するため，放送のような極めて多様な発話環境に適応することは将来にわたる大きな課題である．

本稿では，そのような音響的・言語的な要因に個別的に対応するのではなく，音声認識結果に現れる誤りの性質を利用することで認識率の改善を目指す．すなわち，音声認識をシステムとしてみたときの出力の特性に着目し，どのような誤りが生じやすいのかといった認識誤りの大局的な傾向を統計的にモデル化することを行う．

[1] NHK 放送技術研究所
NHK Science and Technology Research Laboratories, Setagaya, Tokyo 157–8510, Japan
[2] 豊橋技術科学大学
Toyohashi University of Technology, Toyohashi, Aichi, 441-8580, Japan
[a)] kobayashi.a-fs@nhk.or.jp

仮説の誤り傾向は，音声認識結果を学習データとしたときの，単語仮説の正解/誤りのパターンの分布である．したがって，仮説の正誤のパターンを識別的に学習して誤り傾向を反映したペナルティスコアを得れば，認識率を改善できると考えられる．単語仮説の誤り傾向を識別的に学習する方法は，これまで，文献 [3], [4] などによる手法が提案されてきた．文献 [3] では，条件付き対数尤度最大化に基づく学習方法で，文献 [4] では，仮説の期待単語誤り数を直接的に削減するような学習方法でそれぞれ識別的な学習が行われている．これらは，音声データに正解の書き起こしの付与されたラベルありデータに対する教師あり学習の研究であるが，正解の付与されてないラベルなしデータについても，同様の教師なし学習方法が文献 [5] で報告されている．一方，筆者らは，ラベルありデータとラベルなしデータを併用した識別的言語モデルの半教師あり学習の枠組みを提案している [6], [7]．この枠組みは，ラベルあり・ラベルなしのそれぞれのデータに対して Bayes リスク [8] と同様のリスクに基づく学習方法を定義し，多目的最適化問題 (Multi-objective Optimization Programming, MOP)[9] に基づいてこれらを統合する手法である．しかし，文献 [6], [7] では，識別的言語モデルで採用される単語・音素に基づく素性の比較や，ラベルありデータに対するラベルなしデータの量に応じた半教師ありモデルの誤り削減効果については詳細な検討を行わなかった．そこで，本稿では学習データ量や素性関数などの条件を変えた識別的言語モデルによる評価を行い，提案するモデルの有効性に関する議論を行う．

まず，リスク最小化に基づく識別的言語モデルと，多目的最適化に基づく半教師あり学習の方法について述べる．次に，識別的言語モデルで用いる単語・音素に基づく素性関数を説明する．実験では，リスク最小化に基づく識別的言語モデルを用いて放送番組を評価する．そして，条件付き尤度最大化による従来のモデルとの比較や，学習データ量や素性関数などの条件を変えたモデルによる評価から，提案するリスク最小化に基づく識別的言語モデルの有効性を明らかにする．

2. 識別的言語モデル

2.1 対数線形モデル

識別的言語モデルは，入力音響特徴量 $\boldsymbol{x}$ と文仮説 $\boldsymbol{w}$ に対する，次の対数線形モデルとして表される．

$$P(\boldsymbol{w}|\boldsymbol{x};\Lambda) \propto \exp\left\{ f_{\rm am}(\boldsymbol{x}|\boldsymbol{w}) + \lambda_{\rm lm} f_{\rm lm}(\boldsymbol{w}) + \sum_i \lambda_i f_i(\boldsymbol{w}) \right\} \quad (1)$$

ここで，$f_{\rm am}(\boldsymbol{x}|\boldsymbol{w})$ は音響モデルによる対数尤度，$f_{\rm lm}(\boldsymbol{w})$ は言語モデルによる文仮説の対数生成確率とする．$f_i(\boldsymbol{w})$ は素性関数で，特定の単語列もしくは音素列により活性化する (後述)．また，$\lambda_i \in \Lambda$ は素性関数に対する重みであり，識別的言語モデルの推定すべきパラメータとなる．

2.2 Bayes リスク最小化

Bayes リスク最小化は，誤り確率が最小となる仮説を N-best リストから得る上で広く行われている手法である [8]．

$$\hat{\boldsymbol{w}} = \arg\min_{\boldsymbol{w}} \sum_{\boldsymbol{w}'} R(\boldsymbol{w},\boldsymbol{w}') P(\boldsymbol{w}'|\boldsymbol{x}) \quad (2)$$

ここで，$P(\boldsymbol{w}'|\boldsymbol{x})$ は，入力音声 $\boldsymbol{x}$ が与えられたときの，N-best リスト中の文仮説 $\boldsymbol{w}'$ に対する事後確率，$R(\boldsymbol{w},\boldsymbol{w}')$ は $\boldsymbol{w}$ に対する $\boldsymbol{w}'$ の誤り数を表す関数で，一般に 2 つの仮説どうしの Levenshtein 距離 (編集距離) として定義される．本稿で述べるリスク最小化に基づく識別的言語モデルでは，Bayes リスク最小化問題をラティスに適用し，正解ラベルの付与されたラベルありデータ，あるいは付与されていないラベルなしデータから式 (1) のモデルパラメータ Λ を学習する．

2.3 リスク最小化に基づく識別的言語モデル

リスク最小化に基づく学習は，正解ラベルの付与に応じて教師あり，教師なしで行うことができる [4], [5]．識別的言語モデルの教師あり学習は，リスク最小化学習の特別なケースであるため，以下ではラベルなしデータに対するリスク最小化学習について述べる．

まず，リスク最小化学習の目的関数を定義する．学習データとして，正解ラベルの付与されていない発話 $\boldsymbol{x}_m^{(\rm u)}$ $(m = 1,\ldots,M)$ およびその第 k 番目の文仮説 $\boldsymbol{w}_{m,k}$ が与えられたとき，目的関数 (リスク) は次のようになる．

$$\mathrm{U}(\Lambda) = \frac{1}{M} \sum_m \sum_k P(\boldsymbol{w}_{m,k}|\boldsymbol{x}_m^{(\rm u)};\Lambda) \chi(\boldsymbol{w}_{m,k}) \quad (3)$$

ここで，$P(\boldsymbol{w}_{m,k}|\boldsymbol{x}_m^{(\rm u)};\Lambda)$ は文仮説 $\boldsymbol{w}_{m,k}$ の事後確率で，式 (1) から計算される．上のリスクを最小化することにより，学習データに含まれる単語の誤り傾向を反映した対数線形モデルを推定できる．式 (3) の $\chi(\boldsymbol{w}_{m,k})$ はコスト関数で，

$$\chi(\boldsymbol{w}_{m,k}) = \sum_{k'} R(\boldsymbol{w}_{m,k},\boldsymbol{w}_{m,k'}) P(\boldsymbol{w}_{m,k'}|\boldsymbol{x}_m^{(\rm u)};\Lambda) \quad (4)$$

により与えられる．

教師あり学習の場合は，発話 $\boldsymbol{x}_n^{(\rm l)}$ に対して正解ラベルが一意に与えられるため，リスクは次のようになる．

$$\mathrm{L}(\Lambda) = \frac{1}{N} \sum_n \sum_k P(\boldsymbol{w}_{n,k}|\boldsymbol{x}_n^{(\rm l)};\Lambda) R(\boldsymbol{w}_n^{\rm ref},\boldsymbol{w}_{n,k}) \quad (5)$$

式 (4), (5) では，リスクを計算する際に独立した文仮説を用いて計算しなければならない．しかし，個々の文仮説からリスクを計算するのではなく，ラティスの辺どうしの局所的なコストをボトムアップ的に積算していくことで，近

似的かつ効率的にラティスに対するリスクを計算することが可能となる．ラベルありデータを用いた学習では，音響モデルの識別的学習方法である音素誤り最小化 (Minimum Phone Error, MPE) 学習と同様の手法である [10].

ラティスのリスクを計算するために，まず，辺に関する局所的なコスト関数を定義する．これは，Bayes リスクにおける Levenshtein 距離を近似的に置き換えるものである．入力音声 $\boldsymbol{x}_m$ に対するラティス $\mathcal{L}_m$ が与えられたとする．ラティス中のオーバーラップする 2 つの辺 e と e' について，辺に関するコスト関数 $\ell(e,e')$ を次のように定める [5].

$$\ell(e,e') \equiv \begin{cases} 0 & \text{label}(e) = \text{label}(e') \text{ の場合} \\ 1 & \text{それ以外} \end{cases}$$

label 関数は，e に割り当てられた単語仮説を返す関数とする．辺に基づくリスクは，単語誤りに関する局所的な情報を反映するよう，次のように定める．

$$\zeta(e) \equiv \sum_{e' \in \text{overlap}(e)} \ell(e,e') p(e') \tag{6}$$

overlap 関数は，e とオーバラップする辺の集合を返す．$p(e')$ は辺の事後確率で,

$$p(e) = \frac{1}{\bar{\alpha}} \left\{ \alpha(\sigma(e)) \cdot s(e) \cdot \beta(\tau(e)) \right\} \tag{7}$$

によって与えられる．ただし，$\sigma(e)$ は辺 e の始端，$\tau(e)$ は終端をそれぞれ表す．$\alpha(\sigma(e))$ は $\sigma(e)$ における前向き確率で，$\bar{\alpha}$ はラティスの終端における前向き確率を表す．同様に，$\beta(\tau(e))$ は $\tau(e)$ における後ろ向き確率である．$s(e)$ は，音響モデルの対数スコア $\phi_{\text{am}}(e)$，言語モデルの対数スコア $\phi_{\text{lm}}(e)$ および素性関数の重みから，次のように与えられる．

$$s(e) = \exp\left\{ \lambda_{\text{am}} \phi_{\text{am}}(e) + \lambda_{\text{lm}} \phi_{\text{lm}}(e) + \sum_i \lambda_i \phi_i(e) \right\} \tag{8}$$

λ_{am} と λ_{lm} はそれぞれ音響モデル，言語モデルのスコアに対するスケーリング係数で定数とする．$\phi_i(e)$ は，素性 f_i が辺 e で活性化する場合に 1，それ以外では 0 を返すような関数とする．前向きアルゴリズムを使えば，$\zeta(e)$ と $p(e)$ からラティスの累積リスク $\tilde{\gamma}_m$ が計算できる．最終的に，学習データに対するリスクは $\sum_m \tilde{\gamma}_m / M$ と近似される．

素性重み Λ は，学習ラティスのリスクに対する最小化問題を解くことにより得られる．本稿では，準ニュートン法に基づく手法を使ってこの最小化問題を解く．文献 [4] によれば，辺 e における勾配 $\delta^m_{i,e}$ は近似的に

$$\delta^m_{i,e} = -p(e)(\gamma(e) - \tilde{\gamma}_m)\phi_i(e) \tag{9}$$

と計算される．ただし，$\gamma(e)$ は辺 e を通るすべての仮説によるリスクであり，前向き後ろ向きアルゴリズムから計算できる．したがって，第 m 番目のラティスの λ_i に関する勾配は $\sum_{e \in \mathcal{L}_m} \delta^m_{i,e}$ によって与えられ，全学習データに対する勾配 $\Delta^{(\text{u})}_i$ は

$$\Delta^{(\text{u})}_i = \frac{1}{M} \sum_m \sum_{e \in \mathcal{L}_m} \delta^m_{i,e} \tag{10}$$

と計算される．

式 (6) のコストは，ラベルなしデータのすべての仮説を正解とみなして計算したが，ラティスに重畳された正解単語列のみに対して行えば，ラベルありデータに対する教師あり学習を定式化できる．したがって，式 (5) の目的関数値とその勾配 $\Delta^{(\text{l})}_i$ は同様の手順で近似できる．

2.4 条件付き尤度最大化に基づく識別的言語モデルとエントロピー正則化

条件付き尤度最大化に基づく教師あり識別的言語モデルの学習 [3] では，ラベルありデータに対する正解単語列の負の対数尤度の和を目的関数とする．

$$\text{L}(\Lambda) = -\frac{1}{N} \sum_n \log P(\boldsymbol{w}^{\text{ref}}_n | \boldsymbol{x}^{(\text{l})}_n; \Lambda) \tag{11}$$

前節のリスク最小化手法とは異なり，上の目的関数の最適化は，正解単語列の尤度を最大化する一方で，誤りを含む他の対立する文仮説の尤度の和を減少させることに相当する．

対数尤度最大化に基づく目的関数に，ラベルなしデータの情報を統合するため，文献 [11] で述べられているエントロピー正則化に基づく手法を採用する．ラベルなしデータの m 番目の入力音声 $\boldsymbol{x}^{(\text{u})}_m$ が与えられたときの条件付きエントロピーは，次のように定義される．

$$\text{U}(\Lambda) = -\frac{1}{M} \sum_m \sum_k P(\boldsymbol{w}_{m,k} | \boldsymbol{x}^{(\text{u})}_m; \Lambda) \log P(\boldsymbol{w}_{m,k} | \boldsymbol{x}^{(\text{u})}_m; \Lambda) \tag{12}$$

エントロピー正則化では，仮説の識別に関連する不確実性が，推定される識別的言語モデルにより減少するという仮定を置いている．また，この正則化は，文仮説の対数スコアに対する期待値と見なせるので，式 (12) の目的関数の最小化は，正解の可能性のある仮説のスコアを大きくする一方で，見込みのない仮説のスコアを減少させることを意味する．どちらの目的関数も，準ニュートン法を用いて最小化することができる．本稿では，この 2 つの目的関数よる半教師あり学習を従来法に位置づけ，リスク最小化に基づく識別的言語モデルとの比較を行う．

2.5 多目的最適化を利用した半教師あり識別的言語モデル

半教師あり識別的言語モデルは，ラベルなしデータからの情報を，ラベルありデータで学習した教師あり識別的言語モデルに統合することにより，モデルの頑健さの向上を図るものである．本稿では，多目的最適化 [9] に基づいて，2 つの目的関数を統合することで半教師あり学習を行

う [6], [7]．この手法では，目的関数の間のトレードオフを認めることにより，妥協解 (必ずしも最適化されていない解) の集合を得る．妥協解の集合から，最も好ましい解を最適解として選択するが，識別的言語モデルの学習では，学習データとは別に用意した開発データの単語誤り率を最小化するような素性重み Λ を選ぶことに相当する．本稿では，多目的最適化手法の一つである ε 制約法 [12] により 2 つの目的関数を統合する．ε 制約法は，多目的最適化問題の解法の一つとして用いられる [9]．この手法では，一方の目的関数を不等式制約に変換し，他方の目的関数に関する制約付き最適化問題を解く．すなわち，ラベルありデータに対する目的関数 $\mathrm{L}(\Lambda)$，ラベルなしデータに対する目的関数 $\mathrm{U}(\Lambda)$ について，

$$\Lambda' = \arg\min_{\Lambda} \mathrm{L}(\Lambda) \ \text{ subject to } \ \mathrm{U}(\Lambda) \leq \bar{\mathrm{U}} \qquad (13)$$

を解く．ここで，$\bar{\mathrm{U}}$ はあらかじめ設定される目的関数の上限値であり，

$$\bar{\mathrm{U}} = \alpha \mathrm{U}(\mathbf{0}) \qquad (14)$$

によって与えられる．$\alpha (< 1.0)$ はスケーリング定数で，$\Lambda = \mathbf{0}$ における目的関数の値よりも 5 %から 20 %小さな値になるように設定する．目的関数を入れかえれば，$\mathrm{L}(\Lambda)$ を不等式制約とした最適化問題が同様に定められる．この不等式制約付き最適化問題は，拡張ラグランジュ法により解くことができる [13]．統合した目的関数は，

$$\mathrm{F}(\Lambda) = \mathrm{L}(\Lambda) + \rho \left\langle \frac{\kappa}{2\rho} + \mathrm{U}(\Lambda) - \bar{\mathrm{U}} \right\rangle^2 \qquad (15)$$

により与えられる．ここで，κ はラグランジュ乗数，ρ はペナルティパラメータである．また，$\langle x \rangle$ は，$\max\{x, 0\}$ とする．不等式 (13) が満たされない場合の F の勾配は，

$$\frac{\partial \mathrm{F}(\Lambda)}{\partial \lambda_i} = \Delta_i^{(\mathrm{l})} + 2\rho \left(\frac{\kappa}{2\rho} + \mathrm{U}(\Lambda) - \bar{\mathrm{U}} \right) \Delta_i^{(\mathrm{u})} \qquad (16)$$

により求められる．式 (15) の目的関数は準ニュートン法にしたがって最適化する．また，パラメータ κ と ρ は文献 [13] に基づいて更新する．各目的関数に対して定めた 2 つの最適化問題に対して，スケーリング定数 α を変えながら複数の識別的言語モデルを求め，開発データに対する単語誤り率が最小となる素性重みを最適解とする．

2.6 素性関数

本稿の識別的言語モデルでは，単語・音素に基づく素性関数を採用する．単語に基づく素性関数は，文仮説に含まれる単語列により活性化する関数である．例えば，単語 3 つ組による素性は

$$f_{i=h_1(u_1,u_2,u_3)}(\boldsymbol{w}) = c_{u_1,u_2,u_3}(\boldsymbol{w}) \qquad (17)$$

のように，文仮説 $\boldsymbol{w}$ に含まれる 3 つ組 (u_1, u_2, u_3) の数 c_{u_1,u_2,u_3} を返す関数として定められる．ここで，$h_1(u_1, u_2, u_3)$ は，単語 3 つ組に対して該当する素性関数の番号を返すハッシュ関数である [4]．

一方，識別的言語モデルでは，素性重みの学習のために大量の学習データが必要となる．学習データが少なければ，単語仮説どうしの対立関係（誤り傾向）が十分に学習できないため，モデルの効果は限定的となる．そこで，単語よりも粒度が細かい音素に基づく素性関数を導入し，仮説を構成する音素列により対立関係を表現する [14]．音素に基づく素性は，男女のラベルの付与された音素列に基づいて定義されるため，性差を反映した音響的な素性とみなすことも可能である．音素素性は，文仮説 $\tilde{\boldsymbol{w}}$ に対して音素列 $\boldsymbol{q}$ が含まれている数を返す素性として定義される．ただし，音素素性は単語をまたいだ音素列に対しては活性化しないものとする．音声認識の際に得られるラティスは単語を辺とするグラフとなるが，ラティスにデコード時の木構造辞書をトレースした結果 (音素列) を記録しておけば，単語素性と同様に学習できる．認識する際は，音素素性の素性重みを木構造辞書の葉に埋め込み，探索過程で辞書の葉に達した際に音素素性によるスコアを加算する．葉に埋め込まれるスコアは，木の頂点からたどった経路の音素に与えられた素性重みの和として表現される．

3. 実験

3.1 実験条件

識別的言語モデルの学習では，放送音声をデコードして学習ラティスを作成し，単語列または音素列からなる素性関数を抽出する．学習ラティスと素性関数を用いて，リスク最小化または条件付き尤度最大化の各手法に基づいて識別的言語モデルを学習する．音声認識では，音響・言語モデルに加え，識別的言語モデルを用いて 2 パスデコーダによるリスコアリングを行い，認識結果を出力する．

入力音響特徴量は，12 次元の MFCC と対数パワー，および 1 次と 2 次の回帰係数の計 39 次元とした．音声認識は，性別依存 HMM と bigram 言語モデルによりデコードし，200-best 仮説を trigram 言語モデルと識別的言語モデルによりリスコアリングする．音響モデルは，放送ニュース 650 時間から MPE 学習により学習した．ベースラインの trigram 言語モデルは，放送ニュースの書き起こしやニュース原稿 (計 239M 単語) から学習し，語彙サイズを 100k とした．

評価データは，NHK「クローズアップ現代」を 3 番組とした (表 1)．この番組は，番組キャスター (アンカー) とゲスト話者による対談や，ビデオ素材のナレーションを含んでいる．本実験では，1 番組を開発データとして，残りの 2 番組をオープンなテストデータとして利用した．表のパープレキシティ (PP)，単語誤り率 (WER)，未知語率 (OOV) はベースライン trigram 言語モデルによる評価である．

表 1　評価データ
Table 1　Evaluation data for discriminative language modeling

	発話数	単語数	PP	OOV(%)	WER(%)
開発データ	245	3.5k	125.7	1.5	23.0
テストデータ	551	7.0k	139.4	1.3	22.3

表 2　識別的言語モデルの学習データ
Table 2　Training data for discriminative language modeling

	時間	発話数	単語数
ラベルあり	58.6	26k	697.5k
ラベルなし	344.1	218.6k	2.84M

表 3　学習データのパープレキシティと単語誤り率
Table 3　Perplexities and word error rates for training data

	PP	OOV(%)	WER(%)	GER(%)
ラベルあり	64.0	2.03	22.3	13.2
ラベルなし †	163.2	3.07	30.0	16.9

† サブセットに対する数値

表 2 に識別的言語モデルの学習に用いたラベルあり・ラベルなしデータを示す．ラベルありデータは，評価データと同じ「クローズアップ現代」の対談箇所から集めた放送音声である．ラベルなしデータは，報道番組自動書き起こしシステム [2] を用いて収集した放送音声で，ニュースの特定の話題に関する討論などの対談調の発話や，ナレーションなどの原稿読み上げを含む．表のラベルなしデータの発話数と単語数は，音声認識結果から求めた数字である．また，ラベルなしデータは量が多く，全体の単語誤り率を求めることが困難なため，学習データの一部 (4.5 時間分) に対して誤り率を調べた．このサブセットは，2.17 k 発話 (47.2 k 単語) からなる 5 番組とした．表 3 に，ラベルありデータとラベルなしデータのサブセットに対するパープレキシティ，単語誤り率，グラフ誤り率 (GER) を示す．

識別的言語モデルの目的関数の最適化は，L-BFGS アルゴリズム [15] により行った．リスク最小化学習の繰り返し回数，半教師あり学習のパラメータと最適化の繰り返し回数は，表 1 の開発データで定めた．半教師あり学習では，ラベルあり・ラベルなしの各目的関数を不等式制約とみなし，上限を与える α の値を 0.80 から 0.95 まで変え，最適な解を求めた．また，素性関数は，単語 2 つ組，3 つ組および音素 2 つ組，3 つ組とし，ラベルあり・ラベルなしデータで 5 回以上の頻度となるものを定義した (表 4)．

3.2　実験結果

実験では，リスク最小化学習に基づく教師あり，教師なし，半教師ありの識別的言語モデルを従来法 (条件付き対数尤度最大化) と比較した．表 5 に，評価データに対する単語誤り率を示す [7]．教師あり識別的言語モデルでは，尤度最大化とリスクに基づく手法のいずれもベースラインに対する単語誤り率の削減は小さかった．ラベルありデータが少量だったため，統計的に頑健な素性重みの推定に至らなかったことが原因と考えられる．一方，ラベルなしデータのみからリスク最小化で推定した教師なしモデルは，教師ありモデルよりも単語誤り削減率が大きくなった．テストデータに対する単語誤り率は 21.5 %となり，ベースラインの結果と比較すると 3.6 %の単語誤り削減率となった．ラベルなしデータは，ラベルありデータに比べて 6 倍以上の量であるため，より頑健なモデルが得られたと考えられる．半教師あり学習による評価結果では，教師あり，教師なしモデルの結果よりも単語誤り率が改善した．テストデータの単語誤り率は 20.9 %となり，単語誤り削減率はベースラインに対して 6.3 %，教師なしモデル (リスク最小化) に対して 2.8 %となった (危険率 5 %で有意差あり)．

表 4　識別的言語モデルの素性関数
Table 4　Feature functions for discriminative language modeling

		素性数
音素列	2 つ組	1.3k
	3 つ組	12.9k
単語列	2 つ組	731.9k
	3 つ組	1859.6k

4.　考察

4.1　従来法との比較

表 5 の実験結果を見ると，半教師あり学習による識別的言語モデルは，リスク最小化あるいは条件付き尤度+エントロピー正則化のいずれの手法で学習しても，それぞれの教師あり・教師なしモデルの結果に比べて，追加的な単語誤りの削減効果を得ている．半教師あり学習により，ラベルなしデータからの情報が教師ありモデルに効果的に反映されたといえる．一方，リスク最小化に基づく半教師ありモデルは，条件付き尤度最大化+エントロピー正則化に基づくモデルに対して，開発データ，テストデータとも有意に単語誤り率を削減しており (危険率 5 %)，単語誤り削減効果は，リスク最小化学習の方が大きい．

従来法との差異を詳細に調べるため，ラベルありデータに対するラベルなしデータの量を変化させて半教師あり識別的言語モデルを学習して音声認識実験を行った．ラベルなしデータからは，総量の 10，30，50 %に当たるデータをランダムに選択し，識別的言語モデルの学習回数などは，

表 5　識別的言語モデルの実験結果 (単語誤り率，%)
Table 5　Experimental results for discriminative language modeling (WER ,%)

学習方法		開発	テスト
ベースライン (識別言語モデルなし)		23.0	22.3
教師あり	尤度最大化	22.9	22.1
(ラベルありデータのみ)	リスク最小化	22.8	22.3
教師なし	エントロピー正則化	22.7	22.2
(ラベルなしデータのみ)	リスク最小化	22.3	21.5
半教師あり	尤度最大化+エントロピー	22.5	22.0
(ラベルあり＋ラベルなし)	リスク最小化	21.9	20.9

表 6　ラベルなしデータ量を変えたときの半教師あり学習 (単語誤り率, %)
Table 6　Semi-supervised discriminative language modeling with various amounts of unlabeled training data (WER, %)

ラベルなしデータの量	リスク最小化	尤度最大化+エントロピー
10 %	21.3	22.3
30 %	21.2	22.3
50 %	21.1	22.3
100 %	20.9	22.0

表 7　ラベルなしデータ量を変えたときの半教師あり学習 (リスク最小化，%)
Table 7　Semi-supervised discriminative language modeling with various amounts of unlabeled training data (risk minimization, %)

ラベルなしデータの量	対談	VTR
10 %	15.8	27.8
30 %	15.9	27.4
50 %	15.8	27.2
100 %	15.7	27.0
ベースライン	16.4	29.2
教師あり	16.0	29.6
教師なし (100 %)	16.2	27.7

ラベルなしデータをすべて使った際の半教師あり学習の条件と同一とした．表 6 にテストデータに対する単語誤り率を示す．結果を見ると，リスク最小化に基づく識別的言語モデルでは，ラベルなしデータの量が増加するにしたがって単語誤り率が削減されており，ラベルなしデータからの情報が識別的言語モデルに有効に反映されていることが分かる．従来法は，ラベルなしデータの量に関わらず単語誤り率はほとんど変化しておらず，ラベルなしデータの情報がモデルに反映されていない．エントロピー正則化に基づく目的関数は，その定義から，単語誤りに関する情報に依存するのではなく，学習データ中の単語仮説の分布に依存する．したがって，目的関数を最適化することが必ずしも有効な単語誤り率の削減につながるとはいえない．このため，条件付き尤度最大化に基づく目的関数と統合しても，誤りの削減が小幅にとどまったと考えられる．以上の結果から，リスク最小化に基づく 2 つの目的関数を統合した方が，誤り削減に有効な半教師あり識別的言語モデルが得られるといえる．

なお，リスク最小化学習は，ラベルなしデータの量が少量の場合にも大幅に単語誤り率を削減している．学習データが少量であっても，頻度の高い機能語を含む素性の重みの値は統計的な信頼度に関わらず大きな値を取るため，このような素性が単語誤り率の削減に影響を与えたと考えられる．

4.2　リスク最小化学習による半教師ありモデルの効果

リスク最小化学習よる半教師あり識別的言語モデルの効果について，テストデータを「対談 (3.8 k 単語)」と「VTR(ナレーション等, 3.3 k 単語)」の 2 つに分けて評価した (表 7)．表中，対談箇所はアンカーと記者・ゲスト話者による対談で，VTR はナレーター・記者による原稿の読み上げを含む．比較のためベースラインおよびリスク最小化に基づく教師あり・教師なしモデルの結果も併記する．教師あり識別的言語モデルは，ベースラインや教師なしモデルの結果に比べて対談箇所の単語誤りを削減しているが，VTR 箇所はほとんど誤りを削減していない．一方で，教師なしモデルでは VTR 箇所の誤りを大きく削減しており，ベースラインの単語誤り率 29.2 %に対して 27.7 %となり，誤り削減率は 5.1 %となった．これは，ラベルあり・ラベルなしデータの構成の違いが原因ではないかと考えられる．ラベルありデータは同じ番組の対談箇所のみで構成されているが，ラベルなしデータは対談だけではなく，読み上げなどの様々な発話を含んでおり，学習データと評価データとの適合の度合いに応じて識別モデルによる効果の違いが現れたとみられる．また，ラベルなしデータの量が多いことも，教師なしモデルでの単語誤り率の削減に効果があった原因と考えられる．半教師ありモデルの結果を見ると，対談箇所の誤りの削減よりも VTR 箇所の誤り削減の方が大きくなっている．ラベルなしデータをすべて使ったモデルとベースラインの結果を比較すると，対談箇所の単語誤り削減率は 4.3 %(16.4 %から 15.7 %)，VTR 箇所の削減

表 8　リスク最小化学習における素性関数の比較 (%)
Table 8　Comparison of feature functions (%)

	単語	音素	単語＋音素
教師あり	22.2	22.6	22.3
教師なし	21.5	22.4	21.5
半教師あり	21.5	21.8	20.9

率は 7.5 %(29.2 %から 27.0 %) であった．ラベルありデータと大量のラベルなしデータを併用したことにより，大量のラベルなしデータに含まれる読み上げ箇所の誤り傾向の推定が頑健に行えたのではないかと考えられる．

4.3　単語素性と音素素性の比較

単語・音素に基づく素性関数による単語誤り削減効果を調べるために，ラベルあり・ラベルなしデータからそれぞれの素性を使って識別的言語モデルを作成し，比較を行った (表 8)．すべての識別的言語モデルはリスク最小化学習で推定し，テストデータに対する単語誤り率を求めた．単語・音素素性をそれぞれ単独で用いたケースと，両者を組み合わせたケースを教師あり学習で比較すると，素性関数の組み合わせによる単語誤り率の改善はみられなかった．ラベルありデータのサイズが小さく，両者が相補的に機能するような学習が行えなかったものと考えられる．教師なしモデルでは，単語素性のみを用いたモデルの単語誤り率が 21.5 %となって誤りを削減している一方，音素素性を単独に用いたケースでは単語誤り率は 22.4 %となり，ベースラインの結果 (22.3 %) よりも単語誤り率が大きくなった．実験結果から，音素素性はラベルなしデータの誤り傾向を十分にモデル化できていないことが分かる．音素素性は単語素性に比べて粒度が細かい上，ラベルなしデータからの学習では正解単語列による仮説の対立関係が明示されないため，仮説間のコンフュージョンが解消されにくくなり，単語誤り率の削減が困難になったと考えられる．半教師ありモデルの結果を見ると，音素素性のみで学習したモデルの単語誤り率は，単語素性のみのモデルよりも大きくなっているが，教師あり・教師なしの結果よりは誤りが削減されている．音素素性を用いる場合，ラベルなしデータのみでの誤り傾向の学習は困難であるが，ラベルありデータと併用することにより仮説のコンフュージョンが解消され，リスク最小化学習の効果が現れたと予想される．また，単語・音素素性の両者を併用したモデルでは，それぞれの素性を単独に用いたモデルよりも単語誤りが大幅に削減されている．半教師あり学習では，十分な量のラベルあり・ラベルなしデータが与えられたため，単語・音素素性が相補的かつ頑健に学習されたのではないかと考えられる．

5.　おわりに

リスク最小化に基づく識別的言語モデルについて述べた．本稿では，言語的な特徴を用いて識別的言語モデルを学習したが，音声認識の誤り傾向を反映した一般的な識別的モデルでは，言語的素性だけではなく，音響的な素性を利用することも考えられる [16]．言語的な特徴と音響的な特徴を併用すれば，さらなる認識性能の改善が期待できる．また，これらの特徴をマルチパス探索の第 1 パスで利用する場合の有効性についても検討していきたい．

参考文献

[1] 本間真一, 小林彰夫, 奥貴裕, 佐藤庄衛, 今井亨, 都木徹: ダイレクト方式とリスピーク方式の音声認識を併用したリアルタイム字幕制作システム, 映像情報メディア学会論文誌, Vol. 63, No. 3, pp. 331–338 (2008).
[2] 小林彰夫, 奥貴裕, 本間真一, 佐藤庄衛, 今井亨: コンテンツ活用のための報道番組自動書き起こしシステム, 信学論, Vol. J93-D, No. 10, pp. 2085–2095 (2010).
[3] Roark, B., Saraclar, M. and Collins, M.: Discriminative n-gram language modeling, Computer Speech and Language, Vol. 21, pp. 373–392 (2007).
[4] 小林彰夫, 奥貴裕, 本間真一, 佐藤庄衛, 今井亨, 都木徹: 単語誤り最小化に基づく識別的リスコアリングによるニュース音声認識, 信学論, Vol. J93-D, No. 5, pp. 598–609 (2010).
[5] Kobayashi, A., Oku, T., Homma, S., Imai, T. and Nakagawa, S.: Lattice-based risk minimization training for unsupervised language model adaptation, Proc. Interspeech, pp. 1453–1456 (2011).
[6] Kobayashi, A., Oku, T., Imai, T. and Nakagawa, S.: Multi-objective optimization for semi-supervised discriminative language modeling, Proc IEEE ICASSP, pp. 4997-5000, (2012).
[7] Kobayashi, A., Oku, T., Imai, T. and Nakagawa, S.: Risk-based semi-supervised discriminative language modeling for broadcast transcription, IEICE Trans. Inf. & Syst., Vol.E95-D, No.11 (2012, in press).
[8] Goel, V. and Byrne, W.: Minimum Bayes-risk automatic speech recognition, Computer Speech and Language, Vol. 14, pp. 115–135 (2000).
[9] Marler, R. T. and Arora, J. S.: Survey of multi-objective optimization methods for engineering, Structural and multidisciplinary optimization, Vol. 26, pp. 369–395 (2004).
[10] Povey, D. and Woodland, P. C.: Minimum phone error and I-smoothing for improved discriminative training, Proc. ICASSP, pp. I–105–108 (2002).
[11] Grandvalet, Y. and Bengio, Y.: Semi-supervised learning by entropy minimization, Advances in neural information processing systems, pp. 529–536 (2005).
[12] Miettinen, K.: Nonlinear multiobjective optimization, Springer, 1999.
[13] Snyman, J.: Practical mathematical optimization, Springer (2005).
[14] 小林彰夫, 奥貴裕, 本間真一, 今井亨, 中川聖一: 識別モデルにおける音素素性の有効性に関する検討, 音講論集 (春季), No. 2-P-35(a) (2011).
[15] Liu, D. and Nocedal, J.: On the limited memory BFGS method for large scale optimization, Mathematical Programming, Vol. 45, No. 3, pp. 503–528 (1989).
[16] Lehr, M. and Shafran, I.: Discriminatively estimated joint acoustic, duration and language model for speech recognition, Proc. ICASSP, pp.5542–5545, (2010).