Confusion Network を用いた CRF による音声認識誤り訂正

中谷良平[†1]　滝口哲也[†2]　有木康雄[†2]

本稿では，音声認識誤りを CRF (Conditional Random Field) によって検出する誤り検出手法に着目し，認識誤りの自動訂正手法を提案する．CRF による誤り検出手法とは，誤り傾向を示す不自然な N-gram や品詞情報などの言語情報を素性として用いて，誤り検出モデルを学習する．本稿では，Confusion Network 上で誤り検出を行うことで，誤り訂正を実現する．さらに素性として長距離文脈情報を追加し，音声認識の誤認識を減らして精度を向上させる．日本語話し言葉コーパスに対する評価実験の結果，提案手法により，単語誤り率が 32.17%から 28.46%に改善した．

Speech Recognition Error Correction Using CRF on Confusion Network

Ryohei Nakatani,[†1] Tetsuya Takiguchi[†2] and Yasuo Ariki[†2]

This paper presents a speech recognition error correction method using Conditional Random Field (CRF). In previous works, speech recognition error detection methods have been discussed, where the error detection was learned using negative evidence in the form of negative weights on N-grams which are rarely seen in natural language text, or parts of speech information in Japanese. In this paper, we accomplish error correction by detecting error on Confusion Network. Then, we propose error correction using context information. On error correction experiment on speech recognition results, we show that our proposed method improves the recognition performance.

†1 神戸大学 工学部 情報知能工学科
Department of Computer Science and Systems Engineering, Kobe University.
†2 神戸大学 自然科学系先端融合研究環
Organization of Advanced Science and Technology, Kobe University.

1. はじめに

音声は，人間にとって最も身近な情報伝達手段のひとつである．人間同士のコミュニケーションの基本的手段であり，とりわけインタラクティブな情報のやり取りを行う上で，音声は最も有用な手段である．また，テレビ，ラジオ，インターネットなど様々なメディアにおいても多くの音声情報がもたらされている．そのため，音声の検索や機器の操作など，計算機が音声を認識し，理解することが期待される．

現在までの音声研究の結果，音声認識は目覚ましい発展を遂げてきた．特に，計算機の発達によって，統計的音声認識の確立や大規模コーパスの整備などが行えるようになり，音声認識性能は飛躍的に向上した．大語彙連続音声認識において，ニュースなどで読み上げられる書き言葉は，単語正解精度で 95%程度の認識が可能である[1)]．また，学会講演音声のような話し言葉でも，85%程度の精度で認識できるようになってきた．

しかし，まだ十分な音声認識精度が得られたわけではない．機械に誤認識した単語列の不自然さを学習させれば，もっと認識精度を改善できると考えられる．

そこで本研究では，音声認識器が出力した結果について，不自然な単語列があれば自動で訂正することを目的とする．本研究では，大きく次の 2 点に注目している．1 点目は CRF による誤り訂正である．CRF は，誤り部分を特徴づける不自然な N-gram だけでなく，品詞情報や信頼度など，様々な素性を自由に使って誤り傾向を学習できるため，より柔軟な誤り訂正が可能になる．2 点目は長距離文脈情報を用いて訂正精度を向上させることである．CRF によって長距離文脈情報を効果的に取り入れ，Confusion Network[2)] 上で誤り訂正を行う．

以降 2 章では，従来の CRF による音声認識誤り検出手法について述べる．3 章では，長距離文脈情報を意味スコアとして算出する方法と，Confusion Network を用いた CRF による音声認識誤り訂正手法について述べる．4 章では，日本語話し言葉コーパスを用いた音声認識精度の評価実験を行い，その有効性を確かめる．最後に 5 章で本研究の成果をまとめ，今後の課題について議論する．

2. CRF による誤り検出手法

2.1 Conditional Random Field (CRF)

Conditional Random Field (CRF)[3] は，主に自然言語処理やバイオインフォマティクスの分野で用いられているグラフ構造を持つ識別モデルである．文などの構造を持つデータ系列を扱い，モデル式は観測データ系列が与えられたときの出力ラベル系列の条件付確率分布という形をとる．ラベルが与えられた学習データ系列によってモデルを学習し，テストデータ系列を入力すると，モデルが推定するラベル系列が出力される．このとき，データ系列内の各データ一つ一つに最適と推定するラベルを割り当てるのではなく，系列全体として最適と推定するラベルを各データに割り当てる．これはモデル学習時にデータ間の関係も学習し，ラベル推定時にデータ間の関係を考慮した上で各データのラベルを推定することで実現する．

同じ目的で使われるモデルとしては，Hidden Markov Model (HMM) や Markov Random Field (MRF) がある．しかし，これらは観測データ系列と出力ラベル系列の同時確率分布で表現される生成モデルであり，観測データの特徴に強い独立性を仮定する必要がある．この仮定のため，大量の特徴を用いることは困難になるという欠点がある．CRF では条件付確率分布でモデル化しているので特徴間の独立性は考慮する必要はなく，大量の特徴を容易に用いることができる．

2.2 CRF による誤り検出モデルの学習

一般的に，大語彙連続音声認識では統計的言語モデルである N-gram モデルを用いる．これは正しく書かれたテキストから学習を行っており，自然な N-gram にに対して学習は行われるが，不自然な N-gram に対しては学習を行っていない．しかし，従来の N-gram では学習データに存在しない不自然な N-gram についても，スムージングにより推定してしまうため，音声認識結果に不自然な日本語が現れることが少なくない．

そこで誤り検出モデルの学習では，学習に音声認識結果と，対応する正解文書を用い，正解部分，誤り部分で出現しやすい特徴を学習する[4]．例えば,「の-よう-は」のような不自然な N-gram が出現すれば,「は」は誤りである可能性が高いということが考えられ，誤り特徴を示す N-gram として学習される．その結果，不自然な N-gram の発生を抑えることができる．また，表層単語の N-gram に限らず，意味スコアや Confusion Network 上の存在確率など，様々な言語情報が誤り傾向学習に有効であると考えられる．本研究では誤り検出モデルを，認識結果に付与された複数の情報から，各単語に対して正解か誤りかのラベルを付与していく系列ラベリング問題と考え，CRF でモデル化する．

CRF では，入力記号列 x に対する出力ラベル列 y の条件付確率分布 $P(y \mid x)$ を次式のように定義する．

$$P(y \mid x) = \frac{1}{Z(x)} \exp(\sum_a \lambda_a f_a(y, x)) \tag{1}$$

ここで f_a は素性，λ_a は素性関数に対する重みとなる．$Z(x)$ は分配関数で，次式で与えられる．

$$Z(x) = \sum_y \exp(\sum_a \lambda_a f_a(y, x)) \tag{2}$$

パラメータ λ_a は，学習データ $(x_i, y_i)(1 \leq i \leq N)$ が与えられたとき，条件付確率分布 (1) の対数尤度,

$$\mathcal{L} = \sum_{i=1}^{N} \log P(y_i \mid x_i) \tag{3}$$

を最大にするように学習される．これは，正解ラベル列のコストと他のすべてのラベル列のコストとの差が最大になるように学習することに相当する．学習は，準ニュートン法である L-BFGS 法[5] によって行われる．

識別は学習によって得られた確率分布関数 $P(y \mid x)$ を用いて，与えられた入力記号列 x に対する最適な出力ラベル列 $\hat{y}$ を求める問題となる．$\hat{y}$ は次式をもとに Viterbi アルゴリズムにより効率的に求めることができる．

$$\hat{y} = \underset{y}{\operatorname{argmax}}\, P(y \mid x) \tag{4}$$

3. Confusion Network を用いた CRF による音声認識誤り訂正

3.1 Confusion Network

Confusion Network とは，音声認識器の内部状態を簡潔かつ高精度なネットワーク構造へ変換したもので，単語誤り最小化に基づいた音声認識における中間結果である．単語ラティスを音響的なクラスタリングにより，リニアな形式に圧縮することで求められる[2]．図 1 は“私達は”という発話を入力したときの Confusion Network である．破線で囲まれたリンクの集合は，時間的な競合候補を表しており，この集合を Confusion Set と呼ぶ.“-”は，その Confusion Set には単語が存在しないことを表現していて，ヌル遷移と呼ばれる．ヌ

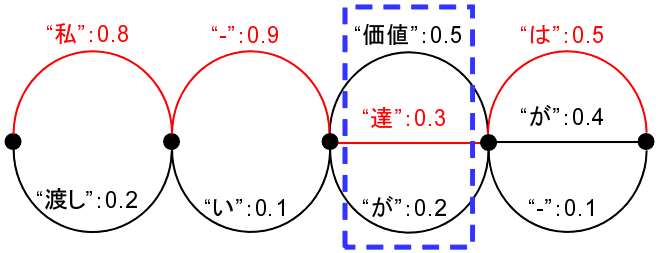

図 1 Example of Confusion Network

ル遷移を選択することでその Confusion Set をスキップすることができる．

また，Confusion Network 上の各単語は，Confusion Set における存在確率 (信頼度) を持っている．Confusion Set はこの存在確率の高い順にソートされるので，第一候補が最も存在確率が高く，以降の競合候補は順に存在確率が低くなっていく．よって，図 1 における第一候補（最上段）を選択していくと，最尤候補が得られる．

3.2 長距離文脈情報

本稿で用いる長距離文脈情報とは，周辺の認識結果単語を参照したときに，識別対象単語の出現が不自然でないかという情報のことである．人間は，N-gram のような部分的な文脈情報だけでなく，より広範囲に渡る長距離文脈情報も考慮しながら音声を聞きとっていると考えられる．例えば図 2 のように,「音声」「会話」「話者」「対話」などの単語が含まれる話題の中に,「大根」という単語が含まれる場合，明らかに不自然である．この存在単語の自然さを意味スコアとして算出し，誤り検出に用いる．しかし，意味スコアは，どの単語と共起しても不自然でない「は」や「です」といった機能語に対しては意味をなさない．そのため，本稿では内容語として名詞，動詞，形容詞のみに意味スコアを与える．

音声認識結果に出現した内容語 w の意味スコア，$SS(w)$ は次のように計算する[4]．

(1) w の周辺に現れる内容語を，図 2 のように文脈窓幅 K で集め，単語集合 $c(w)$ とする (w 自身も含む)．

(2) $c(w)$ 内の各単語 w_i について，$c(w)$ 内の他の単語との類似度 $sim(w_i, c(w))$ を求め，$SC(w_i)$ とする．

$$SC(w_i) = sim(w_i, c(w)) \tag{5}$$

(3) $SC(w_i)$ から，平均 $SC_{avg}(w)$ を求める．

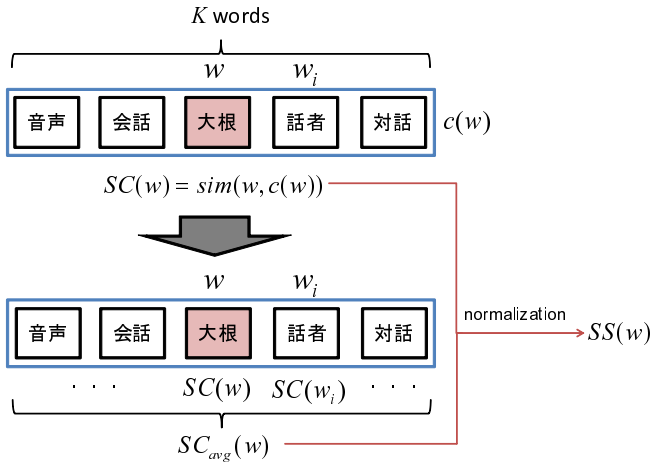

図 2 Calculation of semantic score

$$SC_{avg}(w) = \frac{1}{K}\sum_i SC(w_i) \tag{6}$$

(4) $SC(w)$ と $SC_{avg}(w)$ の差を意味スコア $SS(w)$ とする．

$$SS(w) = SC(w) - SC_{avg}(w) \tag{7}$$

$SC(w)$ が大きいほど周辺に意味が近い単語が多いことになるが，強いトピックを持たない場合，$SC(w)$ は全体的に小さくなってしまう．そのため，SC_{avg} で正規化した $SS(w)$ を意味スコアとして用いる．また，ステップ 2 で出てくる単語間類似度 $sim(w_i, c(w))$ の算出には，潜在的意味解析 (Latent Semantic Analysis : LSA)[6] を用いた．

3.3 CRF による誤り訂正手法

本研究では CRF による誤り検出モデルを利用して，Confusion Network 上で誤り訂正を行う．Confusion Network を用いることで Confusion Set ごとに誤りを訂正できるため，より柔軟な誤り訂正が可能になる．普通，CRF による誤り傾向の学習には音声認識結果の 1-best を用いるが，本研究で用いる Confusion Network には特有のヌル遷移が存在するため，Confusion Network の第一候補単語列（最尤候補），第二候補単語列，第三候補単語列に正誤ラベリングしたものを，CRF によって学習する．ここで，第三候補がない Confusion Set については，第二候補で補い，第二候補がない Confusion Set については，第一候補で

補っている．学習に用いる素性は，次章で述べる．学習後，以下のアルゴリズムに従って誤り訂正を行う．

(1) 評価データを音声認識後，Confusion Network を出力する．
(2) Confusion Network の第一候補列のみを抜き出し，CRF による誤り検出を行う．
(3) 入力時系列順に Confusion Set を見ていく．正解と判定された語には何も操作を行わずに次の Confusion Set へ進む．誤りと判定された語は，対応する Confusion Set から次の候補を選び出し，置き換えてもう一度誤り検出を行う．
(4) Confusion Set の中に正解と思われる語が存在しなければ，存在確率の最も高い語を選択する．
(5) すべての Confusion Set について順番に (3),(4) を繰り返す．

このアルゴリズムの結果，CRF により誤りと判定された語が，正解と判定された語で訂正される．

また、「入力時系列順に」と述べたのは，CRF によって学習する際の素性として bigram, trigram を用いていることから，前の単語が訂正されると，後ろの単語の正誤判定が変わることがあるためである．例えば，2 連続で誤りラベルが付けられている単語列について，1 つ目の単語が訂正されると，bigram 特徴から，2 つ目の単語も正解ラベルに変わることがある．

4. 評価実験

本章では，提案手法の有効性を検証するため，日本語話し言葉コーパス (CSJ) を用いて評価実験を行う．

4.1 実験方法

音声認識器の認識結果に対して，提案手法を用いて誤り訂正を行う．Confusion Network の最尤候補列が，単語誤り率で 29.62% である認識結果を用いる．この認識結果に対して，提案した誤り訂正を行い，単語誤り率がどの程度改善できるかを評価する．また，意味スコアを用いずに誤り訂正を行った結果，Confusion Network の正解単語を全て選んだときの結果，それぞれと比較する．

4.2 実験条件

本研究ではベースとなる音声認識システムに，大語彙連続音声認識エンジン Julius-4.1.4[7] を用いる．以下，システムに必要な音響モデルと言語モデルについて述べる．

音響モデルは，CSJ の学会講演のうち，953 講演 (男性 787 講演+女性 166 講演)，228 時間分の講演音声から作成した HMM を用いた．音響分析条件と HMM の仕様は表 1 のようになっている．1 状態あたりの混合分布数は 16 としている．サンプリング周波数は 16kHz, 音響特徴量は 12 次元 MFCC と対数パワー，12 次元 MFCC の一次微分を加えた 25 次元である．言語モデルは，CSJ の書き起こし文書のうち，2,596 講演の書き起こし文書から学習した N-gram を用いた．

表 1 Speech analysis conditions and specifications of HMM

Sampling frequency	16 kHz
Acoustic feature	MFCC (12 dim.) +Δ MFCC (12 dim.) +Δ power (total 25 dim.)
Window type	Hamming window
Frame length	25 ms
Frame shift length	10 ms
Acoustic model	Triphone (3,000 states)
The number of mixtures	16
State	5 states and 3 loops

次に意味スコアについて述べる．LSA の学習には，CSJ の書き起こし文書のうち，評価データを含まない 2,672 講演のものを用いた．内容語として名詞，動詞，形容詞のみを扱い，語彙数は 48,371 であった．内容語が 30 語程度出現するごとに区切った区間を文書の単位とし，文書数は 76,767 となった．意味スコアを求める際の単語集合 $c(w)$ は，前後 6 発話ずつの Confusion Network における存在確率最大の単語列に，識別対象単語 w を加えたものとした．

CRF による誤り検出モデルの学習には，単語誤り率 35.67% のときの音声認識結果を用いた．また，学習と評価に用いたデータ数を表 2 に示す．学習には 150 講演分，評価には 13 講演分の音声データをそれぞれ用いた．コーパスは CSJ を用いている．学習には，Julius が出力した Confusion Network を用いた．

次に，誤り傾向を学習するための素性を，表 3 に示す．表層単語 N-gram，Confusion Netwrok 上の存在確率，意味スコアを素性として学習した．識別対象の単語を w_0，そこから n 個前の単語を w_{-n}，n 個後ろの単語を w_n と表記している．

また，意味スコアを使った場合と使わなかった場合の CRF による正解検出性能を表 4 に，誤り検出性能を表 5 にそれぞれ示す．評価値は F 値である．どちらも意味スコアを取り入れた方が，検出精度が上がっているのがわかる．

4.3 評価指数

音声認識精度を比較するために，本研究では評価値として単語誤り率 (Word Error Rate

表 2 The number of data

	Learning	Test
Number of lectures	150	13
Number of speeches	39,808	4,771
Number of words	361,513	39,822

表 3 Features used for error tendency learning

Unigram	w_0
Bigram	w_{-1}/w_0 , w_0/w_1
Trigram	$w_{-2}/w_{-1}/w_0$, $w_{-1}/w_0/w_1$, $w_0/w_1/w_2$
Confidence of Confusion Network	$CN\ score$
Semantic score	SC_0

表 4 Performance of correct detection

	Precision	Recall	F-measure
Nonsemantic	0.9136	0.9550	0.9338
Semantic	0.9153	0.9578	0.9361

表 5 Performance of correct detection

	Precision	Recall	F-measure
Nonsemantic	0.7246	0.5673	0.6364
Semantic	0.7403	0.5757	0.6477

:WER) を用いた．WER の定義を以下に示す．

$$WER = \frac{\text{置換誤り数+削除誤り数+挿入誤り数}}{\text{全単語数}} \times 100(\%) \tag{8}$$

「置換誤り」,「削除誤り」,「挿入誤り」について説明する．例えば，

私　達　は　東京　へ　行く

という音声が入力されたとき，

渡し　達　東京　へ　に　行く

という認識結果が出力されたとする．2 つの単語列を比較すると，文頭の「私」が「渡し」と認識されてしまっている．このような誤りが「置換誤り」である．また,「達」のあとの「は」のように抜け落ちているような誤りが「削除誤り」である．そして,「へ」あとの「に」のように挿入されている誤りが「挿入誤り」である．これらは動的計画法 (DP マッチング) によりカウントされる．

表 6 Evaluation with each error type

	SUB	DEL	INS	COR	WER
CN-oracle	1,855	2,476	831	35,491	12.94
CN-best	7,246	2,141	3,423	30,435	32.17
Nonsemantic	6,531	2,633	2,242	30,658	28.64
Proposed method	6,451	2,631	2,253	30,740	28.46

4.4 実験結果

実験方法に基づく実験結果を表 6 に示す.「SUB」は置換誤り,「DEL」は削除誤り,「INS」は挿入誤りの数をそれぞれ表している.「COR」は正解単語の数,「WER」は単語誤り率である.「CN-oracle」は，Confusion Set において常に正解の単語を選択したときの WER である．ただし，正解がないときはその Confusion Set 中で最も存在確率の高い単語を選んでいるため，ヌル遷移が選択されることで削除誤りが最小にはなっていない.「CN-best」は，誤り訂正前のベースとなる，Confusion Network の最尤候補列の単語誤り率,「Proposed method」は，本研究の提案手法である．また「Nonsemantic」は，提案手法の素性として意味スコアを用いない場合の単語誤り率の改善を表している．

表より，Proposed，Nonsemantic ともにベースとなる CN-best 単語誤り率が改善している．意味スコアを使わない Nonsemantic でも CN-best に比べて WER が 3.53 ポイント低くなったことから，CRF によって適切に誤り訂正が行われていることがわかる．また、意味スコアを追加した提案手法では，さらに 0.18 ポイント改善し，CN-best と比較すると 3.71 ポイント改善した． 提案手法によって正しく訂正された例を示す．

- 正解文：実際 の 発話 に
- CN-best：実際 の あ 発話 に
- Nonsemantic：実際 の 発話 に
- Proposed method：実際 の 発話 に

- 正解文：各 モデル の 概要 を 表 に 示し ます
- CN-best：悪 モデル の 概要 を 表 に 示し ます
- Nonsemantic：各 モデル の 概要 を 表 に 示し ます
- Proposed method：各 モデル の 概要 を 表 に 示し ます

- 正解文：ウェブ は

- CN-best：図 は
- Nonsemantic：不 は
- Proposed method：ウェブ は

- 正解文：え イルカ の 頭部 表面 に 伝搬 し て くる
- CN-best：え イルカ の 東部 表面 に 伝搬 し て くる
- Nonsemantic：え イルカ の 東部 表面 に 伝搬 し て くる
- Proposed method：え イルカ の 頭部 表面 に 伝搬 し て くる

---

1 番目と 2 番目の例では、「の/あ」や「悪/モデル」などの不自然な N-gram を持つ認識結果が，CRF によって正しく訂正されている．3 番目と 4 番目の例は，N-gram だけでは不自然と判断できなかった例であるが，意味スコアを加えることで周辺のトピックを考慮し，「図」や「東部」といった単語が訂正されているのがわかる．

4.5 考　　察

CN-best と Proposed method を比較し，誤りの種類別に見てみると，置換誤りが 795 減っている．また，挿入誤りが 1,171 減少，削除誤りが 490 増大していることから，ややヌル遷移が選ばれやすいモデルになっている可能性がある．今回は Confusion Network を使って学習したため，N-gram の中に大量のヌル遷移が現れ，またヌル遷移が正解となることが多いため，適切な学習が行いにくくなっていると考えられる．そのため，ヌル遷移をスキップして学習するアルゴリズムを取り入れることも検討したい．また，CN-oracle と比べると，まだ多くの正解を見落としていて，中でも訂正できる置換誤りが 4,596 個残っている．しかし，意味スコアを導入したことによる改善は主に置換誤りにおいて現れており，80 個の置換誤りが訂正されている．したがって，より効果的な意味スコアを導入することができれば，多くの置換誤りを訂正できると考えられる．

5. おわりに

本稿では，CRF による誤り検出を利用して，Confusion Network 上の誤りを訂正することで，音声認識精度の改善を行った．誤り傾向学習のために様々な素性を自由に取り入れ，特に LSA による意味スコアを取り入れることによって，長距離文脈を考慮した誤り訂正が可能になった．

日本語話し言葉コーパスによる評価実験の結果，単語誤り率において 3.71 ポイントの改善が見られた．これは意味スコアを用いない場合と比べても，0.18 ポイント改善している．

今後の課題として，CRF による誤り検出精度の改善が考えられる．CRF で学習する際の素性として品詞情報，連体形の後には体言しか現れないといった，活用形-品詞の連鎖情報や，助動詞の後に動詞が現れることはめったにないといった，品詞の bigram，trigram などを用いることも有効であると考えられる．その他に，高精度なパラメータ推定を行う[8]ことや，CRF の改良手法[9]による学習を取り入れることも考えたい．

参　考　文　献

1) 中川 聖一, “音声ディクテーションから音声ドキュメント処理へ,” 日本音響学会講演論文集 (秋), pp.1-4, 2007.
2) Lidia Mangu, Eric Brillx, Andereas Stolcke, “Finding consensus in speech recognition: word error minimization and other applications of confusion networks,” *Computer Speech and Language*, 14, pp.41-49, 2000.
3) J. D. Lafferty, A. McCallum, F. C. N. Pereira, “Conditional random fields: Probabilistic models for segmenting and labeling sequence data,” *Proc.ICML*, pp.282-289, 2001.
4) 松本 智彦, 滝口哲也, 有木康雄, “複数の言語情報を用いた CRF による音声認識誤りの検出,” 電子情報通信学会, 音声研究会, SP2008-127, pp.7-12, 2009.
5) J. Nocedal, “Updating Quasi-Newton Matrices with Limited Storage,” *Mathematics of Computation*, pp.773-782, 1980.
6) Thomas Landauer, Peter W. Foltz, Darrell Laham, “Introduction to Latent Semantic Analysis,” *Discourse Processing*, 25, pp.259-284, 1998.
7) “Julius,” http://julius.sourceforge.jp/
8) Christopher White, Jasha Droppo, Alex Acero, Julian Odell, “MAXIMUM ENTROPY CONFIDENCE ESTIMATION FOR SPEECH RECOGNITION,” *ICASSP*, pp.809-812, 2007.
9) Jian Peng, Liefeng Bo, Jinbo Xu, “Conditional Neural Fields,” *NIPS22*, pp.1419-1427, 2009.