重み付き有限状態トランスデューサーを用いた文字誤り訂正

Japanese Character Error Correction using WFSTs

Graham NEUBIG　　森 信介　　河原 達也
京都大学 情報学研究科

1　まえがき

近年情報がますます電子化されるようになり、多くのデータを計算機へ入力する必要がある。カルテの電子化や書籍の電子化 (Google ブック図書館プロジェクト等) がその例である。この作業は光学的文字認識（OCR）や人手による入力などにより行われる。しかし、OCR 誤りや入力ミスなどにより、多くの入力結果は誤りを含む。これらの誤りは、専門的な文章で多くなる傾向がある。この問題を解決するために、分野適応可能な文字誤り訂正システムが求められている。

いくつかの文字誤り訂正法が既に発表されている。しかし、英語を対象とする訂正法 [5, 11] は文字種の多い日本語に不向きであり、日本語を対象とする訂正法 [2, 8, 10] は置換誤りしか扱うことができない。

本論文では、全ての誤りを対象とした分野適応可能な日本語文字誤り訂正システムを提案する。提案手法では、文字誤りを雑音のある通信路モデルでモデル化し、分野依存の言語モデルと応用依存の混同モデルに分割する。本論文では、まず文字誤り訂正に必要な言語モデルについて考え、4 種類の未知語を個別にモデル化し、未知語の性質を正確に捉える。次に、混同モデルの応用依存性について考察する。課題として OCR 誤りを仮定し、図形的特徴を用いて文字混同モデルを構築する方法を提案する。これにより、誤りを含む文と正解の並列コーパスが不要となる。実装においては、重み付き有限状態トランスデューサー（WFST）[7] でモデルを表現することを提案する。これにより、置換誤りのみならず、融合や分離の誤りに対処することが可能となる。

実験では、文字誤りモデルの有効性を検証するために OCR 誤り訂正システムを実装した。実験の結果、文字正解率 97.2% の高いベースラインに対して 12.5% の認識誤りを訂正することができた。

2　文字誤り訂正のモデル化

文字誤り訂正をモデル化するために、機械翻訳などで活用されている雑音のある通信路モデルを用いる。例えば、文字認識結果の誤り訂正の場合、雑音のある通信路は文字認識器であり、スペル誤り訂正の場合では、雑音のある通信路はタイピストである。雑音のある通信路モデルでは、正解文 W がこの通信路により誤りを含む文 O に歪められたと考える。

この枠組みでは、誤りを含む文 O が与えられた時、事後確率 $P(W|O)$ が最も高い文 W を $\hat{W}$ とし、これを正解の文とする。$P(W|O)$ を直接推定することは困難であるため、以下のようにベイズの法則で式を分解する。さらに $P(O)$ を定数として扱うことで、式 (1) が得られる。

$$\begin{aligned}\hat{W} &= \underset{W}{\operatorname{argmax}} P(W|O) \\ &= \underset{W}{\operatorname{argmax}} \frac{P(O|W)P(W)}{P(O)} \\ &= \underset{W}{\operatorname{argmax}} P(O|W)P(W) \end{aligned} \quad (1)$$

式 (1) は 2 つの確率で構成される。$P(W)$ は言語モデル確率であり、W が日本語としてのもっともらしさを表現する。$P(O|W)$ は混同確率であり、正解が W である場合、O が雑音のある通信路によって発生される確率である。以下では、言語モデルを 3 節で、混同モデルを 4 節で記述する。

3　言語モデル

語彙 (既知語の集合) を定義し、それ以外を未知語記号に置き換えて単語 n-gram モデルを構築する。また、未知語の文字列は未知語モデルによりモデル化する。以下では、単語 n-gram モデルと未知語モデルについて順に説明する。

言語モデルはコーパスから学習され、学習コーパスの分野に強く依存する。即ち、コーパスを変えるだけで分野適応ができるようになるため、分野適応は比較的に簡単である。

3.1　単語単位の言語モデル

単語単位の n-gram モデルは長さ $n-1$ の単語履歴から次の単語を確率的に推測する。それぞれの単語確率の

積を取れば文 W の確率が得られる。

$$P(W) = \prod_i P(w_i|w_{i-n+1}^{i-1})$$

後述する実験では 3-gram モデルを利用し、Kneser-Ney 法 [4] で平滑化した。

3.2 未知語モデル

未知語モデルでは、まず単語長をモデル化し、次に文字列を文字単位の n-gram モデルでモデル化する。後述する実験では Witten-Bell 法 [4] で平滑化された 2-gram モデルを利用した。

単語長 k の確率的モデルとして、平均単語長 λ をパラメーターとするポアソン分布で近似する [3]。

$$P_{pois}(k|\lambda) = \frac{(\lambda-1)^{k-1}}{(k-1)!}e^{-(\lambda-1)}$$

特に未知語の中で頻度が高い数字・ローマ字列・カタカナ列は、それぞれ平均単語長や構成文字が大きく異なるため、それぞれを別のモデルで記述することとし、これに、その他の未知語のタイプを加えた 4 種類の未知語のタイプ（数字 t_n, ローマ字列 t_r, カタカナ列 t_k, その他 t_o）を個別に扱うこととした。以上から、タイプ t の未知語 $w_{unk} = c_1c_2\ldots c_m$ の確率は以下の式で表される。

$$P(w_{unk}|t) = P_{pois}(m|\lambda_t)\prod_{i=1}^{m} P(c_i|c_{i-1}, t)$$

ここで c_0 は単語の開始点を意味する特別文字である。

4 混同モデル

混同モデルは誤りを含む文字列と正しい文字列の間の関係をモデル化する。この混同モデルは誤り訂正の応用に依存する。例えば OCR 結果に含まれる誤り文字は視覚的に似ているのに対し、キーボードタイプの誤りは同音意義語の変換ミスなどが多い。

文全体の混同確率 $P(O|W)$ を直接計算するのは困難であるため、各単語における文字混同確率が独立であると仮定し、各文字混同確率の積で文全体の混同確率を近似する。

$$P(O|W) = \prod_{x_j\in O, x_i\in W} P(x_j|x_i) \qquad (2)$$

x は後述する拡張文字を表す。

4.1 文字混同モデルの学習

文字混同モデルの確率は、統計的機械翻訳の翻訳モデルと同じように誤った文字列と正解文字列からなる並列コーパスから学習することができる。しかし、現代の OCR やキーボードタイプの結果の精度は 95% 以上であり、大きな並列コーパスを用意しても混同確率の学習に必要な誤りデータを十分に得ることはできない。この問題に対処するために、並列コーパスから得た確率を平滑化する手法 [8] や誤り傾向を EM アルゴリズムで取得する手法 [11] などが提案されているが、どちらもかなりの量の誤りを含むデータを要し、分野適応が困難となる。

種類	数	割合	例
置換	441	88.91%	維 → 雄
融合	32	6.45%	cl → d
分離	11	2.21%	が → カ 1
挿入	9	1.81%	→ 口
削除	3	0.60%	. →

表 1 誤りの種類とその割合

本論文では、誤りを含むデータを用いることなくその応用における誤り傾向を利用するだけで文字混同確率を十分に近似できると仮定する。この仮定を検証するために、課題として OCR 結果の自動訂正を仮定し、混同モデルの学習に OCR データを利用しない手法を開発し、実際の OCR 誤り訂正での有効性を調べた。次の項で OCR 誤りの傾向とモデルの構築について述べる。

4.2 OCR 誤りの傾向

OCR 誤りの重要な性質の 1 つは、図形的な類似性である。例えば、「維」を「雄」と誤ったり、「cl」を「d」と誤ったりする例が見られる。従って、文字混同確率の計算に図形的特徴を取り入れることが有効である。本論文では、誤りを表 1 のように 5 種類に分類した[*1] 。評価実験で用いられたコーパスにおける誤り種類の割合は表 1 の通りである。この表から、1 対 1 の関係である置換誤りが大半を占め、2 位と 3 位は 2 対 1 の関係である融合・分離誤りであり、最後に 1 対 0 の関係である挿入・削除誤りがあることが分る。1 対 0 の誤りは稀である上、誤りの中でも特に検出しにくく、誤訂正の弊害があり得るので、評価実験では 1 対 0 の誤りについて対応・非対応のモデルを構築し、精度を比較することにした。

4.3 OCR 用文字混同モデルの構築

OCR 誤り訂正で用いられる文字混同モデルは以下の手順で学習した。

1. **拡張文字の作成：**学習コーパスに出現した文字から全ての 2 文字の組み合わせを作成し、新たな文字として扱う。この 2 文字列と通常の文字を合わせて「拡張文字」(X) と呼ぶ。これで 1 対 2 の分離・融合誤りに対応できる。後述する評価実験では効率のため、2 文字列に用いられる文字をローマ字と特殊

[*1] 2 対 1 の誤りは 1 対 1 の誤りと 1 対 0 の誤りによって表すことが可能であるが、「cl」と「d」などの例から分るように、どちらかの文字が削除 (挿入) されたと考えるのは不適切である。

文字に限定した。

2. **図形的特徴の計算：**全ての拡張文字に対して、拡張外郭方向寄与度特徴量 [9] を計算する。この特徴量は、文字が簡単であればあるほどゼロに近くなるので、ヌル文字の特徴量はゼロ列で近似する。これにより 1 対 0 の挿入・削除誤りに対応することができる。
3. **マハラノビス距離：**各特徴量を分散で正規化し、拡張文字間のマハラノビス距離を計算する。
4. **混同確率の計算：**特徴量の分布はマハラノビス空間において正規分布 φ に近いと仮定し、式 (3) のように $P(x_i|x_j)$ を計算する。正規分布の分散 σ^2 は x_j と最も近い点との距離の定数倍を用いる。この定数を変動させることで $P(x_i|x_j)$ における推定文字正解率が変更でき、ここでは推定文字正解率が 96% になるように設定した。

$$P(x_i|x_j) = \frac{\varphi_{\sigma^2}(d_{mahal}(x_j, x_i))}{\sum_k \varphi_{\sigma^2}(d_{mahal}(x_j, x_k))} \tag{3}$$

5 システム実装

前節までで述べた自動文字誤り訂正システムを実装するために、各部分を重み付き有限状態トランスデューサー (WFST)[7] で表現した。

5.1 重み付き有限状態トランスデューサー

重み付き有限状態トランスデューサー (WFST) は有限オートマトンの拡張であり、各状態遷移は入力・出力・重みを有する。入力列に従って状態遷移を繰り返し、その結果、出力と重みが得られる。詳細は文献 [7] を参照されたい。

状態遷移の重みが確率（通常は負対数確率）に相当する場合、$P(O|W)$ や $P(W)$ などの確率的モデルを WFST で表現することができる。複数の WFST を効率的に合成するアルゴリズムや、1 つの WFST を決定化・最小化するアルゴリズムなどが知られているため、個別にコンポーネントを作成しても容易に 1 つのシステムに組み合わせることができる。

5.2 WFST による文字誤りモデル

WFST を用いて英語の文字誤りをモデル化する方法はすでに提案されている [5]。しかし、この方法は未知語に対応しておらず、分かち書きする英語の性質を用いて多少ヒューリスティックに単語分割を扱っている。本論文では、未知語に対応し分かち書きしない言語に対しても利用可能な以下のうようなモデルを提案する。

言語モデル G: この WFST の構築は [7] と同じである。

辞書モデル D: この WFST は負対数確率を加算せずに辞書中の単語を別々の文字シンボルから単語シンボルに変換する。

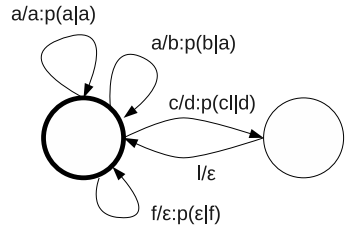

図 1　文字混同モデル WFST

未知語モデル U: この WFST は言語モデルと似ているが、ポアソン分布で確率を与え、未知語 1 語につき 1 つの未知語タグを出力とする。

文字混同モデル T: この WFST は図 1 に示した。図中の「a/a」と「a/b」はそれぞれ単純な 1 対 1 の正解と置換誤りを表現し、「d/c」と「ϵ/l」の列は 1 対 2 の分離誤りを表現し、「f/ϵ」は 1 対 0 の削除誤りを表現している。

誤り訂正システムを構築するために、まず辞書モデル D と未知語モデル U_n、U_r、U_k、U_o の和を取り、文字を単語か未知語タグに変換する L を作る。

$$L = D \cup U_n \cup U_r \cup U_k \cup U_o \tag{4}$$

次に、T と L と G を合成させることで全探索空間 WFST を作成することができるが、T の非決定性により従来の合成法では合成された WFST が巨大になるため、現実的ではない。次節ではこの問題を避ける探索法について述べる。

5.3 探索

全探索空間 WFST を避ける最も単純な方法は、入力文字列から入力 WFST を作成し、T、L、G と逐次合成していく方法である。評価実験で用いた OpenFst[1] ではこの方法のみ可能なため、評価実験ではこの方法を使用した。しかし、これは全探索と同等であり、長い文に対して膨大な記憶容量と計算時間が必要となる。評価実験では、文字混同モデルを枝刈りし探索空間を絞ることで全探索を可能にしたが、この枝刈りによって正解にたどり着けなくなった場合もあり、他の探索方法が望ましい。近年では on-the-fly 合成法が開発されており、ランタイム時の合成においても全探索を回避することが可能となっている [6]。このような探索法を用いることにより、更なる精度向上が期待できる。

6 評価実験と考察

前節までで述べた自動誤り訂正システムの分野適応性能を評価するために、医学分野の大学教科書を評価データとし、学習データとして以下の 3 通りのデータセットを利用した。

データ	1 対 0	F 値	改善率
baseline		97.24	–
train	有	97.36	4.46%
	無	97.37	4.76%
manual	有	97.30	2.21%
	無	97.34	3.79%
merge	有	97.57	12.12%
	無	**97.58**	**12.52%**

表 2　評価実験結果

text:　教科書の第 1 章〜第 13 章の人手による書き起こし（約 44 万字）

manual:　家庭用健康マニュアル（約 3,800 万字）

merge:　text と manual から作成された言語モデルを線形補間により組み合わせたモデル（補間係数は教科書の第 14 章を用いて学習した）

評価データとして、市販のスキャナーおよび OCR による教科書の第 15 章の認識結果と人手による書き起こしを用いた。教科書は高い解像度でスキャンされ、ノイズも少なかったため、文字正解率は 97.2% とかなり高かった。

それぞれのデータセットを用いて、1 対 0 の関係の対応・非対応の場合の自動誤り訂正実験を行った。この結果を表 2 に掲げる。評価基準として、自動訂正の結果と正解文の比較から計算される F 値を用いた。

表 2 から、全ての設定で文字誤り率の改善が見られたことが分る。最も大きい改善が見られたのは 1 対 0 関係非対応の merge であった。merge が大きな改善を見たのは text から文章の書き方についての情報、manual から語彙や文脈についての情報が得られたからであると考えられる。また、1 対 0 対応より 1 対 0 非対応のモデルの方が高い精度となった、訂正すべき 1 対 0 関係が少ない割に誤訂正が多かったからである。

既存の研究では、検出のみを扱っている [2]、あるいは人工的なデータを対象としている [10] などのため、これらとの直接の比較は多少難しい。実際の OCR データを使った研究 [8] では、文字認識器が出した確信度を利用しなければ、97.9% のベースラインに対して約 0.9% の改善しか得ていない。本論文の提案手法は、一般的に考えられる全ての文字誤りを検出・訂正することができるという点と、誤りを含む文と正解文の並列コーパスを必要としない点において、既存の研究に対する優位性がある。

7　むすび

本論文では、文字誤り訂正のための雑音のある通信路モデルを提案し、分野依存の言語モデルと応用依存の文字混同モデルを分離する必要性を主張した。この手法の有効性を実証するため、図形的特徴を用いた OCR 誤り訂正システムを構築した。WFST での実装により全誤り種類に対処でき、個別の未知語モデルにより未知語に対応し、文字の図形的特徴から計算された混同モデルを用いることで分野適応が可能となった。

本論文で提案した文字誤り訂正システムは、文字混同モデルのみを変えることによってスペル誤り訂正に対処できる。将来の研究テーマの 1 つとしてスペル誤り訂正の文字混同モデルの開発とその有効性の検証がある。

参考文献

[1] C. Allauzen, M. Riley, J. Schalkwyk, W. Skut, and M. Mohri. OpenFst: a general and efficient weighted finite-state transducer library. In *Proc. of the CIAA '07*, pp. 11–23, 2007.

[2] 荒木, 池原, 塚原, 小松. マルコフモデルを用いた OCR からの誤り文字列の訂正効果. 情処研報, 94(63):97–104, 1994.

[3] P. E. Brown, V. J. D. Pietra, R. L. Mercer, S. A. D. Pietra, and J. C. Lai. An estimate of an upper bound for the entropy of english. *Computational Linguistics*, 18:31–40, 1992.

[4] S. F. Chen and J. Goodman. An empirical study of smoothing techniques for language modeling. In *Proc. of the ACL96*, pp. 310–318, San Francisco, 1996.

[5] O. Kolak, W. Byrne, and P. Resnik. A generative probabilistic OCR model for NLP applications. In *Proc. of the NAACL03*, pp. 55–62, Morristown, NJ, USA, 2003.

[6] J. McDonough, E. Stoimenov, and D. Klakow. An algorithm for fast composition of weighted finite-state transducers. In *Proc. of the ASRU '07*, 2007.

[7] M. Mohri. Finite-state transducers in language and speech processing. *Computational Linguistics*, 23(2):269–311, 1997.

[8] 永田. 文字類似度と統計的言語モデルを用いた日本語文字認識誤り訂正法. 情処論, 81(11):2624–2634, 1998.

[9] 坂野, 宮本. 拡張外郭方向寄与度法による手書き文字認識. 信学論, 1995(2):282, 19950327.

[10] 竹内, 松本. 統計的言語モデルを用いた OCR 誤り訂正システムの構築. 情処論, 40(6):2679–2689, 1999.

[11] X. Tong and D. A. Evans. A statistical approach to automatic OCR error correction in context. In *Proc. of the WVLC96*, pp. 88–100, 1996.