重み付き有限状態トランデューサーを用いた文字誤り訂正

研究者
Graham NEUBIG　森信介　河原達也

発表者　江口晃

はじめに

- 背景

  -近年情報が電子化され、多くのデータを計算機へ入力する必要がある。

  -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って行われ、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。

研究目的

- 現在の文字誤り訂正
  - -英語を対象とする訂正法
    ⇒文字種の多い日本語には不向き
  - -現在の日本語を対象とする訂正法
    ⇒置換誤りしか扱うことができない

- 研究目的
  - -全ての誤りを対象とした、分野応用可能な日本語文字誤り訂正システムを提案する。

文字誤り訂正モデル

- 文字誤り訂正のモデル化

  -雑音のある通信路モデルを用いる。雑音のある通信路モデルでは、正解文Wがこの通路により誤りを含む文Oに歪められたと考える。

- 文字誤り訂正モデルの式

$$\hat{W}' = \underset{W}{\operatorname{argmax}} P(W|O) = \underset{W}{\operatorname{argmax}} P(O|W)P(W)$$

  -P(W|O)をベイズの法則で分解し、P(O)を定数として扱う。
  -P(W)は言語モデル確率、P(O|W)は混合確率

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。

-未知語の文字列は未知語モデルによりモデル化する

-言語モデルはコーパスから学習する。コーパスを変えるだけで分野応用ができるようになる。

単語単位の言語モデル

- 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の単位履歴から次の単語を確率的に推測する
  -文Wの確率はそれぞれの単語確率の積を取ることで得られる

- 単語単位の言語モデルの式

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

未知語モデル（１）

- 未知語モデル

  -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。

- 単語長のモデル化

  -単語長kの確率的モデルは、平均単語長λをパラメータとするポアソン分布で近似する。

- 単語長のモデル化の式

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

未知語モデル（2）

- 文字単位のモデル化

  -数字・ローマ字列・カタカナ列は平均単語長や構成が大きく異なるため、それぞれ別のモデルで記述する。

- 文字単位のモデル化の式

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

未知語：$w_{unk}=c_1c_2 \ldots c_m$

タイプ：t （数字t_n、ローマ字t_r、カタカナt_k、その他t_o）

混合モデル

- 混合モデル

  -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

  -各単語における文字混合確率が独立であると仮定し、各文字混合確率の積で文全体の混合確率を近似する。

- 混合モデルの式

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

文字混合モデルの学習

- 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、大きなコーパスを用意しても混合確率の学習に必要な誤りデータを十分得ることはできない。
- 誤り傾向を利用するだけで文字混合確率を十分に近似できると仮定する
- 仮定を検証するために、混合モデルの学習に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。

OCR誤りの傾向

- 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

  -誤りを1対1対応の置換誤り、2対1対応の融合・分離誤り、1対0対応の挿入・削除誤りの5種類に分類した

誤りの種類とその割合

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%	. →

OCR文字混合モデルの構築（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。

  これにより1対2の分離・融合誤りに対応できる。

  (評価実験ではローマ字と特殊文字に限定した)

- 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を計算する。

  (この特徴量は文字が簡単であるほどゼロに近くなる)

  これにより1対0の挿入・削除誤りに対応できる。

OCR文字混合モデルの構築（2）

- 混合確率の計算

  -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。

  -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、以下の式のようにP(x_i|x_j)を計算する。

$$P(x_i \mid x_j) = \frac{\varphi_{\sigma^2}(d_{mahal}(x_i, x_j))}{\sum_k \varphi_{\sigma^2}(d_{mahal}(x_i, x_j))}$$

重み付き有限トランデューサー

- 重み付き有限トランデューサー(WFST)は有限オートマトンの拡張であり、各状態変化は入力・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。
- 本実験では状態遷移の重みが確率に相当する

誤りシステムの構築(1)

- 言語モデル　G
- 辞書モデル　D

  -WFSTは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
- 未知語モデル　U
- 文字混合モデル　T

誤りシステムの構築（２）

-辞書モデルDと未知語モデルUn、Ur、Uk、Uoの和を取り、文字を単語か未知語タグに変化するLを作る。

-入力文字列から入力WFSTを作成し、TとLとGを逐次合成していく。

評価実験(1)

- 学習データ
  -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3800万字)
  -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
  -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

評価実験(2)

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%

評価実験(3)

- 考察
  - 全ての設定で文字誤り率の改善が見られた
  - 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

まとめ

-本論文で提案した文字誤り訂正システムは、文字混合モデルを変えることによってスペル誤り訂正に対処できる。

-今後の研究として、スペル誤り訂正の文字混合モデルの開発とその有効性の検証がある。