文献紹介ゼミ

林 秀治

紹介する文献

- Generalized Character-Level Spelling Error Correction
- Noura Farra, Nadi Tomeh, Alla Rozovskaya, Nizar Habash
- Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers), p161-167,

概要

- 文字レベルの変換を対象に、スペルエラーの訂正のための識別モデルを提示
- 文字レベルの動作だが、モデルは語レベルと文脈情報を利用

- 言語非依存の手法だが、今回はEgyptian Arabicの方言テキストで誤りを訂正するためのモデルを適用

背景

- スペルエラー訂正は長年のNLPにおける問題で、それは最近のWebなどのオンラインで生成された未編集の大量のテキストを扱う上で重要
- テキスト内のスペルミスは要約や翻訳のようなタスクでエラーを引き起こすことがある

- 教師あり学習を使用して一般化文字レベルスペルエラー訂正(GSEC)モデルを提示

手法の特徴

- 文字レベル
  - 訂正は教師ありのsequence labelingを使って文字レベルで学習
- 一般化
  - 入力空間はすべての文字で構成され、単一の分類器は全てのトレーニングデータを介して一般的なエラーパターンを学習するために使用
- 文脈依存
  - 文字レベルでの決定を行う際のモデルは単語の文脈を超えて見る

手法の特徴

- 識別
  - モデルは言語固有でも固有でなくてもよく、特徴の数を自由に設定できる
- 言語非依存
  - 言語に依存しないため、形態学的、言語的特徴を考慮していない

今回は他の手法との比較のためにEgyptian Arabicのエラーを修正するためのモデルCODAを適用

文字レベルのスペル訂正

- スペル修正の問題をラベル付与問題として扱う
- 正しい文字を取得するためにそれを変換する方法を記述したaction labelを推測
- action labelの推測にはSVMを使用

文字レベルのスペル訂正

- 入力は、おそらく誤りを含むn文字のe=e1,・・・,enで与えられる
- 今回は、複数の挿入や、挿入および置換の組み合わせなどの複雑な処理が必要な場合(4%)は無視

スペル訂正の手順

- 各文字eiを入力として、誤りが修正されたm文字の文cを以下の操作で生成

  OK:eiは変換されない

  substitute-whith(c)：訓練データを使ってeiを文字cに置換する

  delete:eiを削除

  insert(c):eiの前にcを挿入

スペル訂正の例

入力”korectd”

k:subsitute-with(c)

o:ok

r:insert(r)

e:ok

c:ok

t:ok

d:delete

結果は”correct”になる

使用するデータ

- アラビアの方言には正書法がないので、同じ単語でも複数のスペルがある
- 今回はEskanderらの研究のスペル訂正のための正解テキストを正解とする
- コーパスはEskanderらが使用したものと同じARZコーパスを使用

Features

- 各入力文字が特徴ベクトルによって表される
- Eskanderらの特徴を基に、さらなる改善のために特徴を追加している

  basic features

  - 与えられた文字、前後2文字、語の最初と最後の2文字

  Ngram features

  - 入力文字と共起する前後2~4文字
  - これを加えることで文脈を超えて見ることができる

最尤法(MLE)

- Word-level maximum likelihood modelに基づいて、別のアプローチを実装
- 各入力語をUnigram modelを使い、訓練データに基づいて正しい単語である可能性が最も高い単語に置き換える
- 結果は訓練データに依存

モデルの評価

- 3つの評価尺度を使う
  - Word-error-rate(WER)

    出力の単語レベルの置換、挿入、削除誤りの総数の合計を単語の数で割る
  - Correct-rate(Corr)

    正しい出力単語の数を単語の総数で割る
  - accuracy(ACC)

    正しい出力単語を入力した語の数で割る

Character-level Model Evaluation

- Generalized spelling correction model(GSEC)と比較のためにEskanderらのcharacter-edit classification model(CEC)の性能を示す
- 一つの分類器を使った時、GSECがCECより優れていることがわかる

Approach	Corr%	WER	Acc%
Baseline	75.9	24.2	76.8
CEC	88.7	11.4	90.0
GSEC	89.7	10.4	90.3
GSEC+2grams	90.6	9.5	91.2
GSEC+4grams	91.0	9.2	91.6

モデルの組み合わせ評価

- GSECにMLEを組み合わせる
- GSECの出力に対してMLEを使用

Approach	Corr%	WER	Acc%
MLE	89.7	10.4	90.5
CEC+MLE	90.8	9.4	91.5
GSEC+MLE	91.0	9.2	91.3
GSEC+4grams+MLE	91.7	8.3	92.2

モデルの組み合わせ評価

Approach	Corr%	WER	Acc%
Baseline	75.9	24.2	76.8
CEC	88.7	11.4	90.0
GSEC	89.7	10.4	90.3
GSEC+2grams	90.6	9.5	91.2
GSEC+4grams	91.0	9.2	91.6

Approach	Corr%	WER	Acc%
MLE	89.7	10.4	90.5
CEC+MLE	90.8	9.4	91.5
GSEC+MLE	91.0	9.2	91.3
GSEC+4grams+MLE	91.7	8.3	92.2

実験

- 結果をよりよく理解するためにデータ内の単語を
  - 訓練データにあるまたはin-vocabulary word(IV)
  - out-of-vocabulary(OOV)

  にわける
- MLEはIVカテゴリのほとんどをカバーできるはず
- OOVを見ることでGSECの性能がわかる

実験結果

- IVではCECとGSCは同等の性能
- OOVではGSECのほうが良好である

	Input Words	Baseline	CEC+MLE	GSEC+MLE
OOV	3,289(17.2%)	70.7	76.5	80.5
IV	15,832(82.8%)	78.6	94.6	94.6
Total	19,121(100%)	77.2	91.5	92.2

結果の考察

- IVから50単語のエラーを人手で解析
  - 同じ文字への変換(X→X)
  - CECによってモデル化される変換(X→Y CEC)
  - CECによってモデル化されない変換(X→Y not CEC)
  - 複雑なエラー(complex)

Type	Characters	Example	CEC	GSEC
X-X	16502	m-m,space-space	99.25	99.33
X-Y(CEC)	609	y-ý	80.62	83.09
X-Y(not CEC)	161	T-θ,del{w}	31.68	43.48
Complex	32	n-ins{A}{m}	37.5	15.63

結果の考察

- X-Y(not CEC)ではGSECのほうが優れている
- complexでは、CECが複数の分類器を適用して補正することができるので優れている

Type	Characters	Example	CEC	GSEC
X-X	16502	m-m,space-space	99.25	99.33
X-Y(CEC)	609	y-ý	80.62	83.09
X-Y(not CEC)	161	T-θ,del{w}	31.68	43.48
Complex	32	n-ins{A}{m}	37.5	15.63