文献紹介ゼミ

林 秀治

紹介する文献

- Chinese Spelling Error Detection and Correction Based on Language Model, Pronunciation, and Shape
- Yu Junjie, Li Zhenghua
- Proceedings of The Third CIPS-SIGHAN Joint Conference on Chinese Language Processing (2014)
- pp.220-223

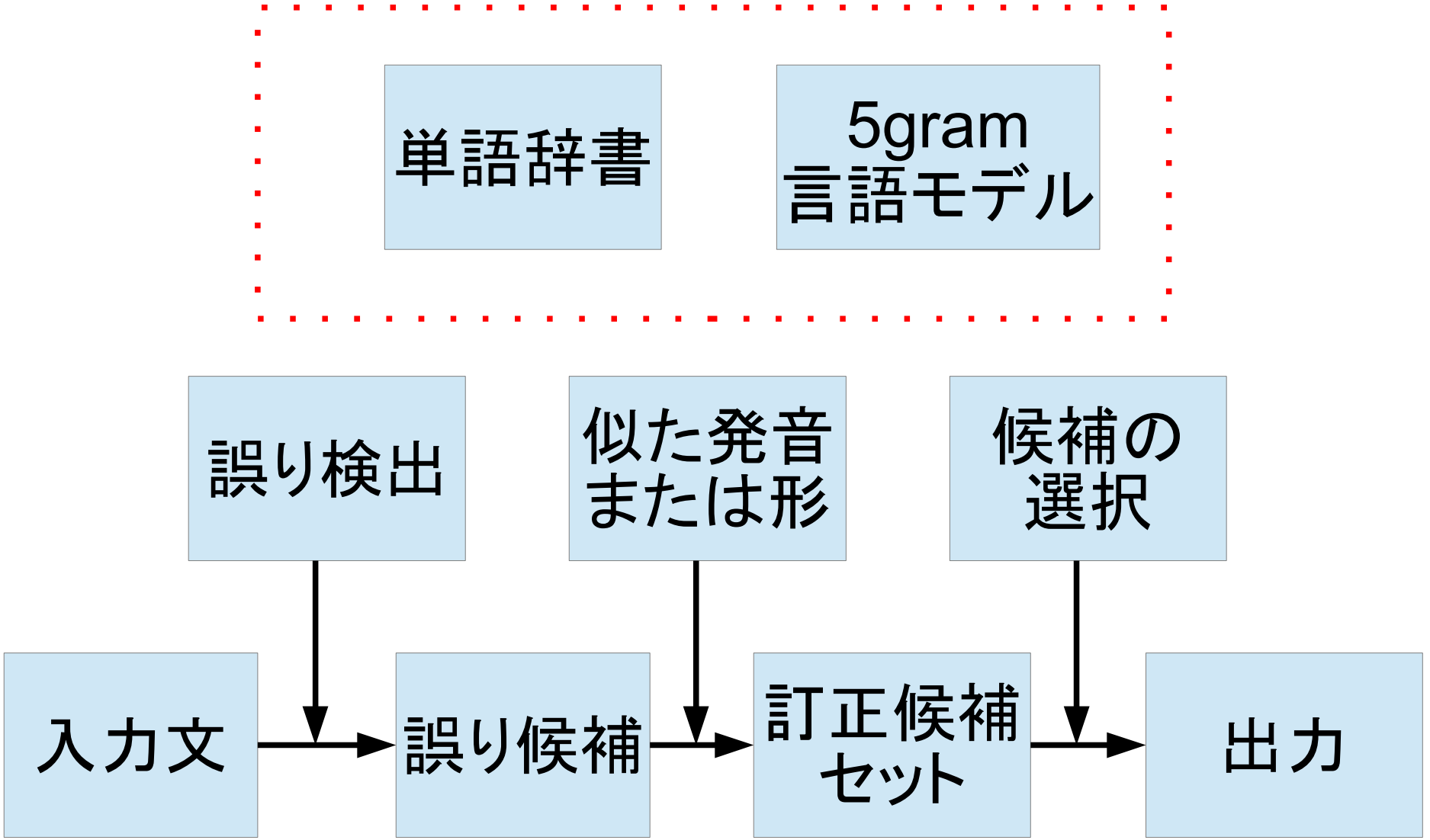
概要

- ngram言語モデルを使って誤りを検出・訂正
- 対象文字の発音・形状に類似したものの候補セットを生成
- 言語モデルの確率が一番高いものを出力
- 精度は高いが再現率が低い（精度重視）

Introduction

- 英語でのスペル誤り検出・訂正では、誤りは非単語と実単語に分類できる
- しかし、中国語では単語分割がされておらず、すべての文字は実単語
- 単語分割せずに文字単位で検出し、単語単位で訂正

システムの概要



Resources

- 約1200万文のChinese Gigaword version 2.0を言語モデルの学習に使用
  - 単語毎に分割せず、文字単位で学習
- 誤りのみ単語レベルで検出できるよう、Webから収集した約30万の単語辞書を構築
- SIGHANが提供している5,000の漢字についての似た発音・形状の辞書を使用

誤り検出

- 二つの方法を提案
  - 言語モデルでスコアが低いもの
  - 自動単語分割後に独立した文字

- 誤りとして検出されたものは正しい文字が多い
- 過去の研究によると、学生の書いた文章に含まれる誤りの数は平均2と少ない

誤り検出

- 2つの手法を組み合わせ候補を絞る

- 5gram言語モデルで文の各文字のスコアを計算
  - スコアが閾値以下のものを候補に
- 前後5文字の範囲を見て単語になるか確認
  - 単語がなければ最終的な誤り候補とする

誤り訂正

- 最初に、誤り訂正のための候補セットを作成
- 似た発音・形状の文字の誤りが多い
  - 似た発音・形状辞書を基に候補を作成

- 候補セットの文字に置換し、隣接文字と単語を形成するか確認
  - 単語を形成できるものは言語モデルでスコアを計算
- 5gram言語モデルで最も高いスコアが閾値より高ければその文字に訂正

誤り訂正の例

- 『竟』が対象文字
- 『逆竟』、『竟時』、『逆竟時』、『到逆竟』、『到逆竟時』、『遇到逆竟』、『遇到逆竟時』が辞書にあるか
- 辞書にない場合、『竟』に似た発音・形状の候補セットを作る
- セットの文字に置換、スコアが最も高いものが閾値より高ければそれに訂正

実験

- SIGHAN organizerの1,000文をテストデータとして使用
- SIGHAN Bake-off 2013の最終テストの300文を訓練データとして使用
- 訓練データは402文字の誤りを含む

閾値による再現率の変化

- 閾値があがるにつれ再現率も上がるが、1文で検出する平均文字数も増える
- 訓練データの平均長さは70文字なので、閾値が-1のときは平均で半数以上が誤り
- 今回閾値は-2に

Function threshold	Language model	
	Recall(%)	#Characters
-4	26.67	2
-3	57.00	6
-2	86.67	18
-1	96.32	38

結果

- 再現率が低いので、F1値も低い
- 誤検出の割合は0.032と非常に低く、検出精度も0.74とそこそこ高い
- しかし、再現率の低さは無視できないので改善が必要

Run	False Positive Rate	Detection Level				Correction Level			
		Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
1	0.2524	0.4539	0.3881	**0.1601**	**0.2267**	0.4426	0.3527	**0.1375**	**0.1978**
2	**0.032**	**0.5292**	**0.7385**	0.0904	0.1611	**0.5235**	**0.7119**	0.0791	0.1424