文献紹介ゼミ

林 秀治

紹介する文献

- Combining Trigram-based and Feature-based Methods for Context-Sensitive Spelling Correction
- Andrew R. Golding, Yves Schabes
- ACL '96 Proceedings of the 34th annual meeting on Association for Computational Linguistics
- pp.71-78

概要

- {peace,piece},{quiet,quite}のようなタイプミスによるエラーや{amout,number}のような使用法のミスを訂正する
- 従来の手法に品詞trigramやベイズを使ったものがあるがそれぞれ長所と短所がある
- そこでtrigramとベイズを組み合わせた手法を提案

Introduction

- 従来のスペルチェッカーは対象語が単語リストにない場合誤りとする
- タイプミスによる誤りの15%は実際にある単語
  - quite→quietなど
- 過去の研究ではこのような語が原因の訂正誤りが全体の25~50%(Kukich, 1992)

Introduction

- このような語を訂正するため文脈依存のスペル訂正を行う
  - 例：Can I have a peace of cake?
  - {peace,piece}のようなsetを使いどちらが正しいか予測
- いくつか手法があるが今回は品詞trigramとベイズ分類器を使った方法に注目
- それぞれ長所と短所があるので2つをうまく組み合わせたTribayesを提案

Trigrams

- 品詞triguramを使用
- 対象語をset内の各単語に置き換えて、それぞれ確率を計算
- set内の語が違う品詞を持っている場合うまく機能するが同じ場合はうまく機能しない

Bayes

- Bayesian hybrids(Golging, 1995)を使用
  - ベイズ分類器を使ったContext wordsとCollocationsの組み合わせ
- set内の単語の品詞が異なる場合結果が悪い
  - 文法により決定されるため、Trigramのほうが結果が良い

TrigramとBayesの比較

Confusion set	Different tags				Same tags			
	Break-down	System scores			Break-down	System scores		
		Base	T	B		Base	T	B
their, there, they're	100	56.8	97.6	94.4	0	—	—	—
than, then	100	63.4	94.9	93.2	0	—	—	—
its, it's	100	91.3	98.1	95.9	0	—	—	—
your, you're	100	89.3	98.9	89.8	0	—	—	—
begin, being	100	93.2	97.3	91.8	0	—	—	—
passed, past	100	68.9	95.9	89.2	0	—	—	—
quiet, quite	100	83.3	95.5	89.4	0	—	—	—
weather, whether	100	86.9	93.4	96.7	0	—	—	—
accept, except	100	70.0	82.0	88.0	0	—	—	—
lead, led	100	46.9	83.7	79.6	0	—	—	—
cite, sight, site	100	64.7	70.6	73.5	0	—	—	—
principal, principle	29	0.0	100.0	70.0	71	83.3	83.3	91.7
raise, rise	8	100.0	100.0	100.0	92	61.1	61.1	72.2
affect, effect	6	100.0	100.0	66.7	94	91.3	93.5	97.8
peace, piece	2	0.0	100.0	100.0	98	44.9	42.9	89.8
country, county	0	—	—	—	100	91.9	91.9	85.5
amount, number	0	—	—	—	100	71.5	73.2	82.9
among, between	0	—	—	—	100	71.5	71.5	75.3

Breakdownは割合
Baselineは頻度

Tribayes

- set内の単語がすべて同じ品詞の場合Bayesを、違う品詞の場合Trigramを適用する

Confusion set	Different tags			Same tags		
	Break-down	System scores		Break-down	System scores	
		T	TB		B	TB
their, there, they're	100	97.6	97.6	0	—	—
than, then	100	94.9	94.9	0	—	—
its, it's	100	98.1	98.1	0	—	—
your, you're	100	98.9	98.9	0	—	—
begin, being	100	97.3	97.3	0	—	—
passed, past	100	95.9	95.9	0	—	—
quiet, quite	100	95.5	95.5	0	—	—
weather, whether	100	93.4	93.4	0	—	—
accept, except	100	82.0	82.0	0	—	—
lead, led	100	83.7	83.7	0	—	—
cite, sight, site	100	70.6	70.6	0	—	—
principal, principle	29	100.0	100.0	71	91.7	83.3
raise, rise	8	100.0	100.0	92	72.2	75.0
affect, effect	6	100.0	100.0	94	97.8	95.7
peace, piece	2	100.0	100.0	98	89.8	89.8
country, county	0	—	—	100	85.5	85.5
amount, number	0	—	—	100	82.9	82.9
among, between	0	—	—	100	75.3	75.3

各手法の比較

Confusion set	System scores			
	Base	T	B	TB
their, there, they're	56.8	97.6	94.4	97.6
than, then	63.4	94.9	93.2	94.9
its, it's	91.3	98.1	95.9	98.1
your, you're	89.3	98.9	89.8	98.9
begin, being	93.2	97.3	91.8	97.3
passed, past	68.9	95.9	89.2	95.9
quiet, quite	83.3	95.5	89.4	95.5
weather, whether	86.9	93.4	96.7	93.4
accept, except	70.0	82.0	88.0	82.0
lead, led	46.9	83.7	79.6	83.7
cite, sight, site	64.7	70.6	73.5	70.6
principal, principle	58.8	88.2	85.3	88.2
raise, rise	64.1	64.1	74.4	76.9
affect, effect	91.8	93.9	95.9	95.9
peace, piece	44.0	44.0	90.0	90.0
country, county	91.9	91.9	85.5	85.5
amount, number	71.5	73.2	82.9	82.9
among, between	71.5	71.5	75.3	75.3

Microsoft Wordとの比較

- Microsoft Word (version 7.0）と比較
- 訂正候補がある場合それを適用
- 訂正後も訂正候補がある場合はそれを適用
- ループする場合は最初の候補を適用
- 指摘だけで候補がない場合はスキップ

Wordの結果との比較

Confusion set	Tribayes		Microsoft Word	
	Correct	Corrupted	Correct	Corrupted
their, there, they're	99.4	87.6	98.8	59.8
than, then	97.9	85.8	100.0	22.2
its, it's	99.5	92.1	96.2	73.0
your, you're	98.9	98.4	98.9	79.1
begin, being	100.0	84.2	100.0 *	0.0 *
passed, past	100.0	92.4	37.8	86.5
quiet, quite	100.0	72.7	100.0 *	0.0 *
weather, whether	100.0	65.6	100.0 *	0.0 *
accept, except	90.0	70.0	74.0	36.0
lead, led	87.8	81.6	100.0 *	0.0 *
cite, sight, site	100.0	35.3	17.6	66.2
principal, principle	94.1	73.5	11.8	94.1
raise, rise	92.3	48.7	92.3	51.3
affect, effect	98.0	93.9	100.0	77.6
peace, piece	96.0	74.0	36.0	88.0
country, county	90.3	80.6	100.0 *	0.0 *
amount, number	91.9	68.3	100.0 *	0.0 *
among, between	88.7	54.8	97.8	0.0