文献紹介ゼミ

林 秀治

紹介する文献

- A Bayesian hybrid method for context-sensitive spelling correction
- Andrew R. Golding
- In Proceedings of the Third Workshop on Very Large Corpora, pages 39-53, Cambridge, MA, 1995

概要

- 2つの方法が文字の曖昧性を解消するのに有効
  - 曖昧な目標単語への一定距離内にある語
  - 対象の単語の周りの単語と品詞

- この2つの手法はカバレッジが違うので、これらを組み合わせて更なる改善を目指す

Context-sensitive spelling correction

- 辞書内の正しい言葉になるようスペルミスを訂正
- 従来のスペルチェッカでは検出されない
  - タイプミス：{out,our}
  - 同音異義語：{there,their}
  - 語法：{between,among}

Context-sensitive spelling correction

- 単語の曖昧さをconfusion setをつかってモデル化
- C={w1,・・・,wn}はセット内の単語wiと曖昧である
- C={desert,dessert}があるとき、どちらかが文章中に現れたとき文脈からどちらが正しいか推測
- 今回セットは先行研究[Flexner,1983]のものを使用

Context-sensitive spelling correction

- 単語の曖昧さをconfusion setをつかってモデル化
- C={w1,・・・,wn}はセット内の単語wiと曖昧である
- C={desert,dessert}があるとき、どちらかが文章中に現れたとき文脈からどちらが正しいか推測
- 今回セットは先行研究[Flexner,1983]のものを使用

5つの訂正手法

- 比較のため5つの手法を用意
- Baseline：ほかとの比較のための『最低限の能力』
- Context words：対象語の±k語
- Collocations：対象語の周りの構文パターン
- Decision lists：リストを使って上記を組み合わせ
- Bayesian cassifiers：ベイズ分類器を使って上記2つを組み合わせ(提案手法)

使用コーパス

- 1-million-word Brown corpus [Ku cera and Francis, 1967]を訓練に使用
- 3/4-million-word corpus of Wall Street Journal text [Marcus et al., 1993]をテストに使用

Baseline

- 曖昧な語が出現した時、トレーニングコーパス中で最も頻出した語とする

Baseline

Confusion set	No. of training cases	No. of test cases	Most frequent word	Baseline
whether, weather	331	245	whether	0.922
I, me	6125	840	I	0.886
its, it's	1951	3575	its	0.863
past, passed	385	397	past	0.861
than, then	2949	1659	than	0.807
being, begin	727	449	being	0.780
effect, affect	228	162	effect	0.741
your, you're	1047	212	your	0.726
number, amount	588	429	number	0.627
council, counsel	82	83	council	0.614
rise, raise	139	301	rise	0.575
between, among	1003	730	between	0.538
led, lead	226	219	led	0.530
except, accept	232	95	except	0.442
peace, piece	310	61	peace	0.393
there, their, they're	5026	2187	there	0.306
principle, principal	184	69	principle	0.290
sight, site, cite	149	44	sight	0.114

Context words

- 対象語の±k語Cjを使って最も可能性の高い語wiを推定する。
- ベイズの定理を使って算出

$$p(w_i|c_{-k},\ldots,c_{-1},c_1,\ldots,c_k) = \frac{p(c_{-k},\ldots,c_{-1},c_1,\ldots,c_k|w_i)\,p(w_i)}{p(c_{-k},\ldots,c_{-1},c_1,\ldots,c_k)}$$

Context words

Confusion set	Baseline	Cwords ±3	Cwords ±6	Cwords ±12	Cwords ±24
whether	0.922	0.902	0.922	0.927	0.922
I	0.886	0.914	0.893	0.883	0.851
its	0.863	0.862	0.795	0.743	0.702
past	0.861	0.861	0.849	0.801	0.743
than	0.807	0.931	0.901	0.896	0.855
being	0.780	0.791	0.795	0.793	0.755
effect	0.741	0.747	0.741	0.759	0.716
your	0.726	0.816	0.783	0.774	0.736
number	0.627	0.646	0.622	0.636	0.639
council	0.614	0.639	0.614	0.602	0.614
rise	0.575	0.575	0.575	0.585	0.498
between	0.538	0.759	0.697	0.671	0.586
led	0.530	0.530	0.530	0.521	0.557
except	0.442	0.695	0.526	0.516	0.558
peace	0.393	0.754	0.705	0.574	0.574
there	0.306	0.726	0.623	0.557	0.466
principle	0.290	0.290	0.290	0.290	0.435
sight	0.114	0.455	0.250	0.364	0.318
Avg no. of context words		27.9	36.9	55.9	92.9

Collocations

- 語と品詞を要素として使用
- {desert,dessert}のとき、～the _という文章を探す
  - Travelers entering from the desert were …
- それぞれの語に対して品詞タグを与える
  - walkなら{NS,V}(singular noun,verb)

Collocations

Confusion set	Baseline	Collocs ≤ 1	Collocs ≤ 2	Collocs ≤ 3
whether	0.922	0.939	0.931	0.931
I	0.886	0.979	0.981	0.980
its	0.863	0.943	0.945	0.950
past	0.861	0.919	0.909	0.909
than	0.807	0.966	0.965	0.966
being	0.780	0.853	0.853	0.842
effect	0.741	0.821	0.821	0.821
your	0.726	0.877	0.887	0.887
number	0.627	0.646	0.646	0.681
council	0.614	0.663	0.639	0.639
rise	0.575	0.807	0.807	0.807
between	0.538	0.699	0.730	0.733
led	0.530	0.849	0.840	0.863
except	0.442	0.800	0.789	0.789
peace	0.393	0.869	0.869	0.852
there	0.306	0.911	0.932	0.932
principle	0.290	0.841	0.812	0.812
sight	0.114	0.341	0.318	0.318
Avg no. of collocations		33.9	263.1	985.4

Decision lists

- Yarowsky [1994]がcontext wordsとcollocationsの相互関係を指摘
- 両方の長所を取得する方法として決定リストを提案
- Contextとcollocationを要素としたリストを作る
- 各要素を信頼性でソート

Decision lists

Confusion set	Baseline	Cwords ± 3	Collocs ≤ 2	Dlist Rely	Dlist $U(x\|y)$
whether	0.922	0.902	0.931	0.935	0.829
I	0.886	0.914	0.981	0.980	0.808
its	0.863	0.862	0.945	0.931	0.805
past	0.861	0.861	0.909	0.932	0.892
than	0.807	0.931	0.965	0.967	0.961
being	0.780	0.791	0.853	0.842	0.933
effect	0.741	0.747	0.821	0.821	0.654
your	0.726	0.816	0.887	0.868	0.896
number	0.627	0.646	0.646	0.629	0.667
council	0.614	0.639	0.639	0.627	0.651
rise	0.575	0.575	0.807	0.804	0.827
between	0.538	0.759	0.730	0.659	0.800
led	0.530	0.530	0.840	0.840	0.840
except	0.442	0.695	0.789	0.789	0.726
peace	0.393	0.754	0.869	0.852	0.836
there	0.306	0.726	0.932	0.914	0.906
principle	0.290	0.290	0.812	0.812	0.841
sight	0.114	0.455	0.318	0.432	0.568

Bayesian cassifiers

- 取得可能なすべての証拠を考慮することでさらなる改善が規定できると仮定
- リスト内の各要素を照合して対象語を分類

結果の比較

Confusion set	Baseline	Cwords ±3	Collocs ≤ 2	Dlist Rely	Bayes Rely	Trigrams
whether	0.922	0.902	0.931	0.935	0.935	0.873
I	0.886	0.914	0.981	0.980	0.985	0.985
its	0.863	0.862	0.945	0.931	0.942	0.965
past	0.861	0.861	0.909	0.932	0.924	0.955
than	0.807	0.931	0.965	0.967	0.973	0.780
being	0.780	0.791	0.853	0.842	0.869	0.978
effect	0.741	0.747	0.821	0.821	0.827	0.975
your	0.726	0.816	0.887	0.868	0.901	0.958
number	0.627	0.646	0.646	0.629	0.662	0.636
council	0.614	0.639	0.639	0.627	0.639	0.651
rise	0.575	0.575	0.807	0.804	0.807	0.574
between	0.538	0.759	0.730	0.659	0.786	0.538
led	0.530	0.530	0.840	0.840	0.840	0.909
except	0.442	0.695	0.789	0.789	0.811	0.695
peace	0.393	0.754	0.869	0.852	0.852	0.393
there	0.306	0.726	0.932	0.914	0.916	0.961
principle	0.290	0.290	0.812	0.812	0.812	0.609
sight	0.114	0.455	0.318	0.432	0.455	0.250