文献紹介ゼミ

林 秀治

紹介する文献

- Word Vector/Conditional Random Field-based Chinese Spelling Error Detection for SIGHAN-2015 Evaluation
- Yih-Ru Wang and Yuan-Fu Liao
- Proceedings of the Eighth SIGHAN Workshop on Chinese Language Processing (SIGHAN-8) (2015)
- pp.46-49

概要

- 外国人学習者が書いた中国語の文章のスペルミスの検出
- 単語ベクトル/CRF(Conditional Random Fields)ベースの検出器を提案

背景

- 中国語のスペルチェックは異常な単語列の検出問題として扱うことができた
- 多くの場合、言語モデル(LM)を使って検出が行われてきた
- 検出対象は中国人の書いた中国語の文であった

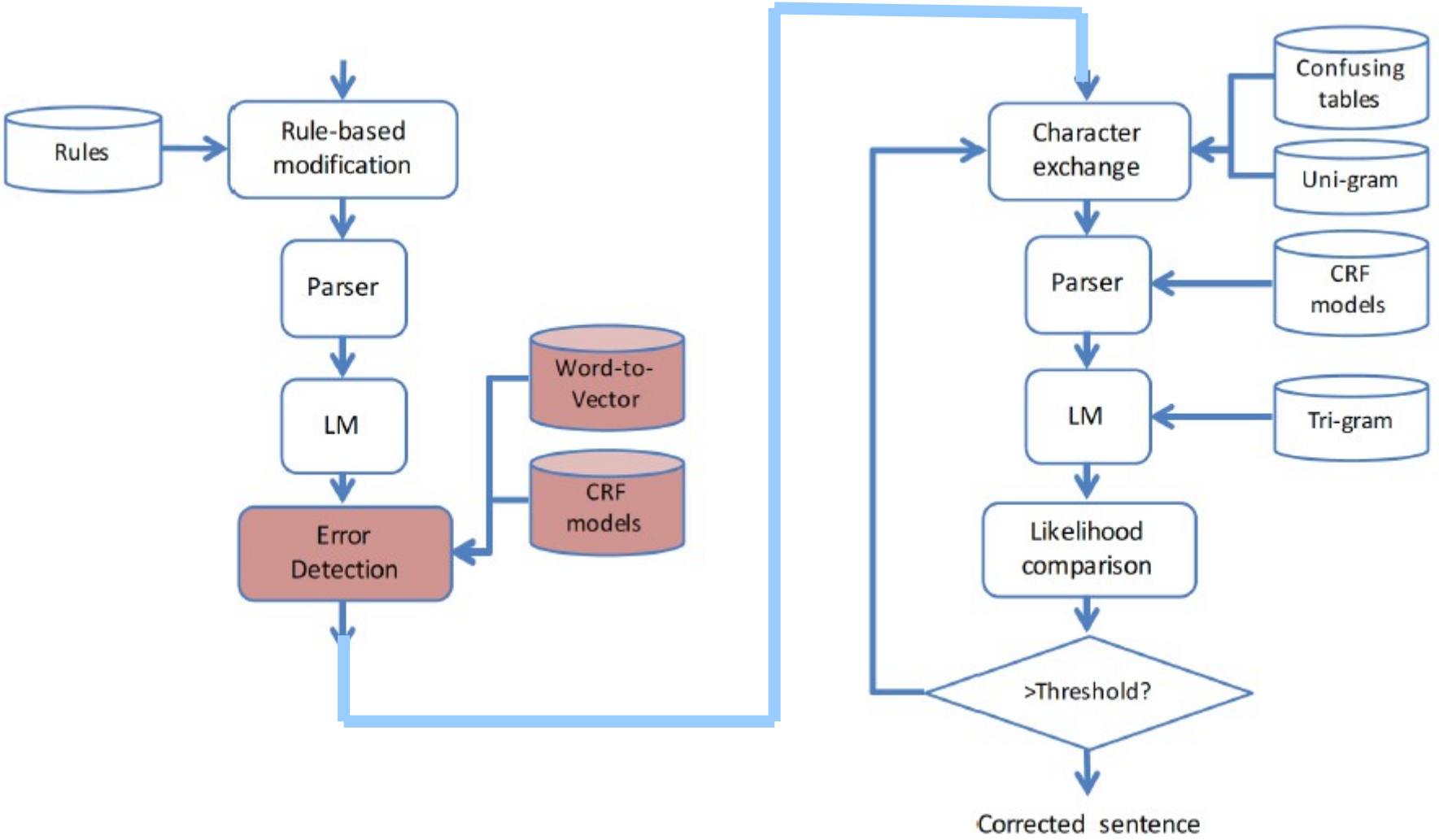
- 外国人学習者の書いた中国語では従来の手法は有用ではないかもしれない

背景

- 学習者は文法的な間違いのほかにも文法はあっているが間違った意味の語を使うことがある
- このタイプはLMを使って検出することが困難

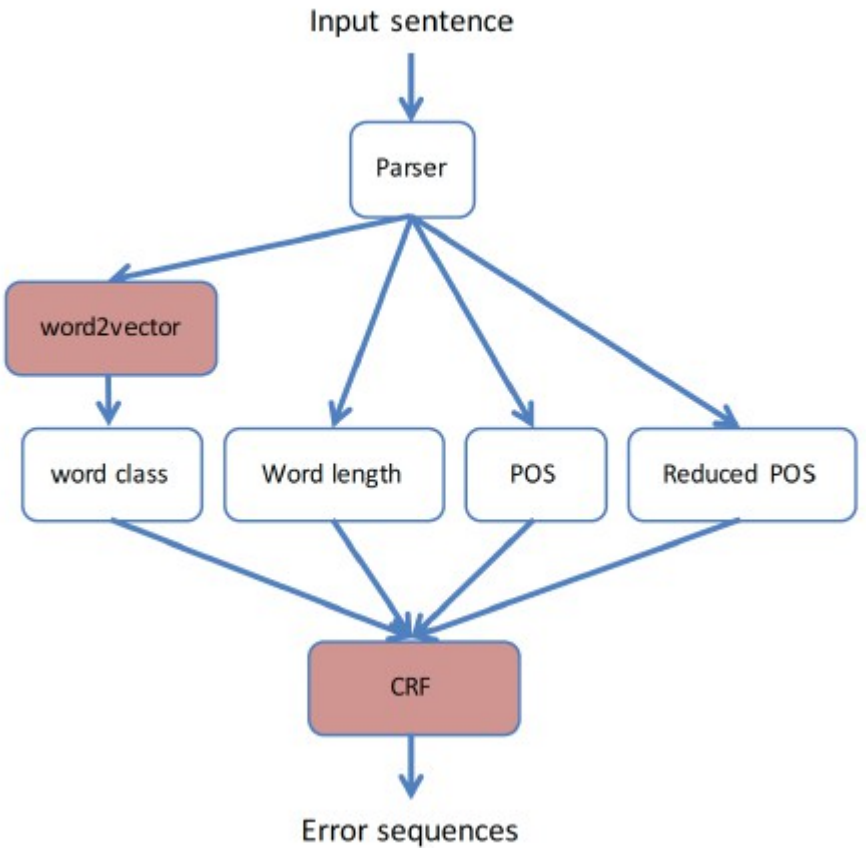
- 単語ベクトルを構築し、CRFベースの手法で誤りを検出

システムの概要


Rules
Rule-based modification
Parser
LM
Word-to-Vector
CRF models
Error Detection
Character exchange
Confusing tables
Uni-gram
Parser
CRF models
LM
Tri-gram
Likelihood comparison
>Threshold?
Corrected sentence

Word Vector/CRF-based Spelling Error Detector

- word2vecとCRFを使って誤り検出を行う



CRF Chinese spelling error detector

- Word vectorとParserの出力を1つのfeatureに結合する
- CRFは正誤ラベルのサンプルのセットから学習

Example of training sample

- 各単語は単語の長さ、品詞、品詞タグ、word class index、正誤ラベルの５次元ベクトルに変換

聽起來	3	D	ADV	436	c
是	1	SHI	Vt	441	c
一	1	Neu	DET	136	c
份	1	Nf	M	162	c
很	1	Dfa	ADV	441	c
好	1	VH	Vi	398	c
的	1	DE1	T	390	c
公司	2	Nc	N	609	c
。	1	PM	M	-2	c
又	1	Caa	C	551	w
意思	2	Na	N	77	c
又	1	Caa	C	551	c
很多	2	Neqa	DET	441	c
錢	1	Na	N	270	c
。	1	PM	PM	-2	c

実験

- parser、tri-gram LM、Word vectorはSinica Balanced Corpus version 4.0（約44億語）で学習
- parserの単語分割のF値は、オリジナルで96.72%

  人手で修正したコーパスで97.67%
- 47種の品詞タグ付けは精度94.24%
- Word vectorの窓は17(8+1+8)語を使用

実験

- また、CRFベースの誤り検出器を作るために、Bake-off2014とSIGHAN2015のコーパスを使用(合計で106,815語、内4.537語が誤り)
- テストセットは11,808語で498の誤りを含んでいる

Frontend result

- Vector/CRF-based spelling error detectorの出力
- 0.5以上は正、0.5未満は誤り

但是	2	Cbb	C	441	0.9999
我	1	Nh	N	738	0.9998
不能	2	D	ADV	441	0.9833
去	1	D	ADV	738	0.9945
參加	2	VC	Vt	723	0.9985
，	1	PM	PM	-2	0.9998
因為	2	Cbb	C	441	0.9999
我	1	Nh	N	738	0.9999
有一點	3	Dfa	ADV	738	0.9997
事情	2	Na	N	441	0.9687
阿	1	T	T	820	0.0048
！	1	PM	PM	-2	0.9999

Evaluation result

- Bake-off2014、SIGHAN-2015コーパスの誤り検出の結果
- 訓練データの検出結果は非常に良い
- しかし、単語の正誤で数に差があるので問題

		Acc.	Pre.	Rec.	F1
Training	C		99.92	99.98	99.95
	W		99.21	97.47	98.33
	All	99.90	99.90	99.90	99.90
Test	C		98.23	99.03	98.63
	W		54.10	38.98	45.31
	All	97.32	97.32	97.32	97.32

Overall result

- 閾値0.999、0.98、0.95のCRFの3システム(Run1~3)を使用
- SIGHAN2015での結果

Run	F/P	Detection Level			
		Acc.	Pre.	Rec.	F1
1	0.050	0.605	0.837	0.261	0.398
2	0.065	0.609	0.812	0.283	0.420
3	0.132	0.601	0.717	0.336	0.457
Run	F/P	Correction Level			
		Acc.	Pre.	Rec.	F1
1	0.050	0.578	0.802	0.207	0.329
2	0.065	0.580	0.776	0.227	0.351
3	0.132	0.564	0.663	0.261	0.375