N-gram 誤り判定器を複数用いた 中国語学習者のための誤り検出

関　磊　　指導教師：山村　毅

1.　はじめに

近年，中国の世界での存在感が増えており，IT など様々な分野で中国の影響力が広がっている。それとともに中国語学習者もだんだん増えている。しかし，中国教育機関の推定では現在全世界で約 500 万人の中国語教師が必要となると言われている。「人間の教師が足りないなら，プログラムの力を貸したらどう？」「少なくともプログラムで文法誤り検出ぐらいできる？」というような考えが出現するのは自然なことであろう。したがって本研究では，中国語学習者を書いた文章を対象に，中国語文の誤りを自動検出システムを開発する。

2.　関連研究

近年，中国語誤り自動検出に関しては，機械翻訳，CRF(条件付き確率場)モデル，LSTM(Long short-term memory)などの手法を利用する研究が多く行われている。

中国語学習者を書いた中国語の誤り検出を共有課題として，世界中の各チームを集めて，同じデータに対する誤り検出を競争しながら，最新技術を交流する研究会（NLP-TEA）がある。NLP-TEA2018 の最新結果では，すべてチームの正解率の平均値は 0.5599 であった。すべてのチーム中 HFL チームは最優正解率 0.7278，F 値 0.7283 であった。そして，CMMC チームは最優 F 値 0.7563，正解率 0.6889 であった [1]。

3.　本研究の特徴

自然言語処理の分野では，文章誤りを検出したい場合は，言語モデルを利用するのが一般的である。先行研究 [2] では，中国語生コーパスから中国語単語 bi-gram 確率モデルを生成し，単語 bi-gram 確率モデルの有効性を検証したが，偶発ゼロ頻度問題が生じていた。本研究ではこの手法について改良を行い，4 つの手法を提案し，これに基づき，9 種類 n-gram 誤り判定器を作成した。さらに，生成した 9 種類の n-gram 誤り判定器を組み合わせることでアンサンブル誤り判定器を提案する。

4.　N-gram 誤り判定器

表 1　n-gram 誤り判定器出力例

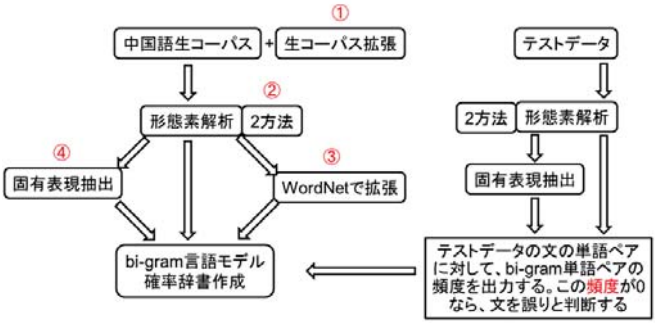
訳文：	私は今日はとても暑いと思う
正解文：	我 觉得 今天 很热
実行例：	(我,觉得):11 (觉得,今天):15 (今天,很热):4
誤り文：	我 今天 很热 觉得
実行例：	(我,今天):2 (今天,很热):4 (很热,觉得):0

n-gram 誤り判定器の基本の検出手法は学習データから正しい中国語についての bi-gram モデルを作成し，テストデータの単語 bi-gram ペアに対して，表 1 に示すのような単語 bi-gram ペアの頻度を出力するものである（カッコの後の数字が頻度を表す）。この頻度が 0 なら，文を誤りと判断する。

5.　偶発的ゼロ頻度問題への対応

偶発的ゼロ頻度問題の対策として四つの手法を提案している。一つ目は生コーパスの拡張である。学習データ（129 万点の新聞記事，合計 1.76GB の SogouCA コーパス）に加えて 70 万点の記事，合計 3GB の weixin コーパスを利用するものである。二つ目は形態素解析ツール自身の 40 万単語辞書の代わりに NLPCN の 360 万単語辞書を利用するものである。三つ目は中国語 WordNet（ハルビン工業大学情報検索研究室類義語辞書：類義語，同類語は合計 13440 組 72966 単語）を利用し，n-gram 言語モデルを拡張するものである。人名，地名など固有表現により，偶発 0 頻度が生じったことがあるので，四つ目は固有表現を抽出し（Stanford NER を利用），タグに置き換えてから，n-gram 言語モデルを作成するものである。これらの手法を組み合わせて，n -gram 誤り判定器合計 9 種類を作成した。全体像を図 1 に示す。

図 1　n-gram 誤り判定器作成手順



偶発ゼロ頻度に対応した例を表 2 に示す。形態素解析ツールの辞書を変えることで，単語の句切りが変わるため，誤検出がなくなっている。WordNet で拡張したり，生コーパスを拡張するのも有効である。また，地名の固有表現抽出することで，bi-gram モデルのサイズを大幅で減らしながら，ゼロ頻度問題にも対応することができる。

表 2　各手法対応例

入力した正解文	法国的交通很方便
誤検出結果	('法国的', '交通'):0('交通', '很方便'):9
訳文	(フランスの,交通)(交通,とても便利)
生コーパス拡張	('法国的', '交通'):1('交通', '很方便'):47
形態素解析辞書の変換	('法国', '的'):4773('的', '交通'):12513('交通', '很'):155('很', '方便'):3263
WordNet 拡張	('法国的', '交通'):6('交通', '很方便'):17
固有表現抽出	('LOC', '交通'):1478('交通', '很方便'):9

6. N-gram による誤り検出実験

実験の対象となるデータにはTOCFL作文コーパスの4721文を用いた。これは正誤文の割合が大体半々で与えられているものである。9種類n-gram誤り判定器の実験結果を表3に示す。

表 3　9種類 n-gram 誤り判定器の実験結果

	生コーパス	手法	偽陽性率 (FPR)	正解率 Acc	適合率 Pre	再現率 Rec	F値
形態素解析：40万単語辞書	①		26.64%	67.17%	73.36%	56.95%	64.12%
	①+②	生コーパス拡張	20.05%	63.48%	79.95%	38.86%	52.30%
	①	WordNet 拡張	18.52%	60.73%	81.48%	30.76%	44.66%
	①	固有表現 PER	31.53%	65.71%	68.47%	61.97%	65.06%
形態素解析：360万単語辞書	①		39.43%	64.88%	60.57%	91.16%	72.78%
	①+②	生コーパス拡張	33.25%	68.44%	66.75%	77.18%	71.59%
	①	WordNet 拡張	35.77%	66.19%	64.23%	77.59%	70.28%
	①	固有表現 PER+LOC +ORG	28.36%	58.31%	71.64%	31.58%	43.84%
	①	文字tri-gram	28.77%	67.06%	71.23%	60.49%	65.42%

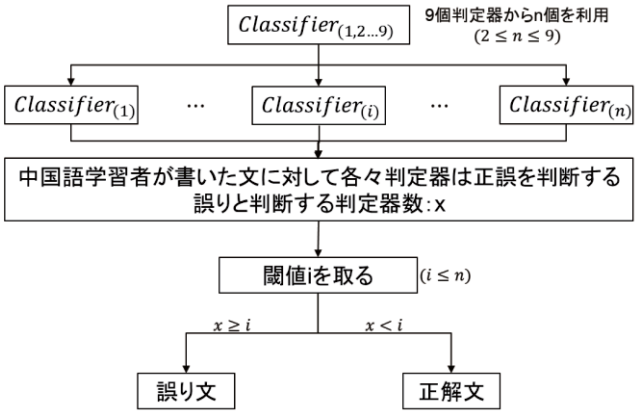
正解率が一番高いのは，形態素解析ツール Jieba で 360 万単語辞書を使い，SogouCA コーパスと Weixin コーパス両方を利用した場合である。適合率が一番高いのも形態素解析ツール Jieba で 40 万単語辞書を使い，SogouCA コーパスに対して，中国語 WordNet を利用し，単語 bi-gram 辞書を拡張した場合である。再現率と F 値が一番高いのは，形態素解析ツール Jieba で 360 万単語辞書を使い，SogouCA コーパスだけを用いた場合である。比較として文字 tri-gram 誤り判定器も作成し，単語 bi-gram 誤り判定器と比較すると，単語 bi-gram の成績がちょっと高いである。

形態素解析の効果について，360 万単語辞書のほうが再現率と F 値が高いという傾向が分かる。適合率から見ると 360 万単語辞書を使った場合は誤判断したことが多いことが言える。生コーパスを拡張した場合と WordNet で拡張した場合は形態素解析方法に関係なく，どちらにおいても適合率が上がった。反対に再現率は下げている。ゼロ頻度問題に効果があることが分かる。固有表現抽出の性能について、一つ固有表現特徴「人名」(PER) を使った場合は F 値が高くなった。しかし，三つの固有表現特徴「人名」,「地名」,「組織名」を使った場合は再現率が低くなった。その原因は StanfordNER の精度の問題である。固有表現ではない単語を固有表現と判断してしまったためと考えられる。

7. アンサンブル誤り判定器の提案手法

アンサンブル誤り判定器における，文の正誤の判断基準は，まず表 3 にあげた 9 種類の n-gram 誤り判定器 Classifier1…Classifier9 中から n(1<n≤9)個の判定器を選んで，文の正誤を判断する。つぎにこの文に対して，誤りと判断された判定器数を x とするとき，閾値 i(1≤i≤n)を設けて，もし x≥i なら，この文は誤りと判断する。つまり，n 個判定器中で i 個以上の判定器が誤りであると判断されたら，この文は誤りと判断する。その流れを図 2 に示す。

図 2　アンサンブル誤り判定器による誤り検出



8. アンサンブル誤り判定器の実験結果

n-gram 誤り判定器の選択と閾値の選択は重要である。アンこの実験では，すべての判定器の組み合わせ合計 2295 種類を生成し，その性能を評価した。そのなかで正解率が一番高いのは固有表現（PER，LOC，ORG）以外 8 種 n-gram 誤り判定器を利用し，閾値を 4 にした場合である。その結果を表 4 に示す。F 値が一番高いのは 9 種類 n-gram 誤り判定器を利用し，閾値を 4 にした場合である。その結果を表 5 に示す。

表 4　最高正解率アンサンブル誤り判定器

用いた分類器番号	Fpr	Acc	Pre	Rec	f1	閾値
(1, 2, 3, 4, 5, 6, 7, 9)	26.30%	75.53%	75.74%	77.26%	76.49%	4

表 5　最高 F 値アンサンブル誤り判定器

用いた分類器番号	Fpr	Acc	Pre	Rec	f1	閾値
(1, 2, 3, 4, 5, 6, 7, 8, 9)	30.36%	75.15%	73.76%	80.35%	76.91%	4

9. まとめ

表 3 により WordNet や固有表現を利用することで，コーパスを増やすことと同様の効果が得られることが分かった。偽陽性率を下げ，適合率を上げ，0 頻度問題に効果があるといえる。固有表現抽出ツールと形態素解析ツールによる実験結果を影響することが大きいことも分かった。

本研究では，アンサンブル誤り判定器の手法を提案し，有効性を検証した。N-gram 誤り判定器より，全面的な性能上昇することができる。N-gram 誤り判定器の最高成績より正解率 6.71 ポイント，F 値 5.32 ポイント良く，NLP-TEA2018[1] の最優秀成績と同等レベルな結果が得られた。

文　献

1. 「Overview of NLPTEA-2018 Share Task Chinese Grammatical Error Diagnosis」, Gaoqi Rao Qi Gong Baolin Zhang Endong Xun, Beijing Language and Culture University, Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications, pages 42–51
2. 「中国語学習者の誤り支援」, 関磊, 山村毅, 電気・電子・情報関係学会東海支部連合大会講演論文集, C1-7, 名古屋大学, 2017
3. 「統計的機械翻訳を用いた中国語文法誤り訂正」, 趙 寅琛,小町守,石川 博,研究報告自然言語処理（NL）,2016-NL-225(6),1-6
4. 「A Hybrid Model for Chinese Spelling Check」,Hai Zhao,Deng Cai,Yang Xin,Yuzhu Wang,Zhongye Jia, ACM Transactions on Asian and Low-Resource Language Information Processing, Volume 16 Issue 3, April 2017 Article No. 21.