大規模学習者コーパスを用いた語学学習者支援

小町研究室（自然言語処理）准教授 小町守 / 協力: 奈良先端大 水本智也, 澤井悠, 坂口慶祐; Microsoft Research 荒瀬由紀

概要

ウェブからマイニングした大規模な学習者コーパスは、専門家（語学教師）がタグ付けした小規模なコーパスと比較して、質は低くても様々なタスクで有用である。

① 語学学習SNS Lang-8 の添削ログから大規模な語学学習者コーパスを抽出した
② 前置詞・冠詞・動詞選択などの誤りの自動訂正や問題生成が高精度に行なえることを示した

研究背景

背景
✓語学学習者の作文の誤りを自動訂正したい　He works ~~to~~at ~~the~~a flowershop.

問題点
①一般的に入手可能な語学学習者の作文データが少ない（数千文）
②語学学習者の誤りのパターンを考慮していない（例: 母語の影響）

目的
①大規模な語学学習者コーパスの構築
②語学学習者の誤りパターンを用いた作文誤り訂正・問題生成

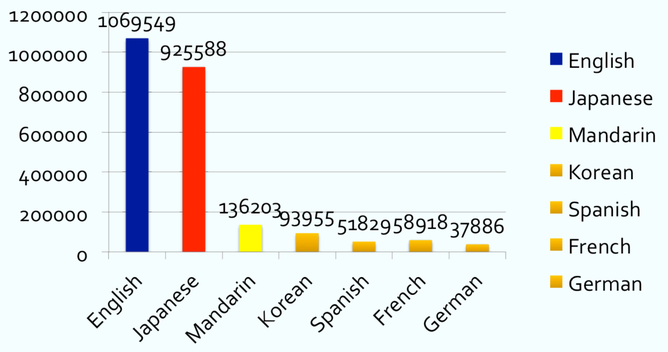
応用
✓英作文の誤りの自動訂正エンジン
✓誤用の例文の検索エンジン
✓選択肢問題の自動作成・採点

提案手法

添削文対からの大規模言語学習者コーパス構築

Sentence written by a JSL learner	三人はそれぞれ自分 の方式で 感情を表 れ ます。
Sentence corrected by an annotator1	三人はそれぞれ自分 なりの表現で 感情を表 し ます。 Each of three expresses their feelings in their way of expression.
Sentence corrected by an annotator2	三人はそれぞれ自分 なりに 感情を表 し ます。 Each of three expresses their feelings in their own way.

Lang-8: 言語交流 SNS http://lang-8.com/
- ユーザ数: 210,834 (2010年11月)
- ユーザが日記を投稿し、1文ごとにネイティブスピーカーが添削する。1つの文が複数の添削を受けることもある。



利点
✓ 大規模に獲得可能
✓ 複数の添削候補が取得できる

欠点
✓ 誤りの種類が分からない
✓ クラウドソースされた添削の品質は一定ではない

コーパスの量を変えた実験

✓ 冠詞、前置詞、語彙選択誤り訂正はコーパスの量が増えると性能が向上した
✓ 名詞の数、時制、主語・動詞の一致の誤り訂正はコーパスの量を増やしても効果がない

Training Corpus	KJ	Lang-8						
		2K	10K	20K	100K	200K	300K	390K
article	0.277	0.282	*0.390	*0.420	*0.443	*0.459	*0.475	*0.488
noun number	0.308	0.226	0.214	0.238	0.270	0.300	0.319	0.311
preposition	0.201	0.143	0.192	0.226	*0.333	*0.336	*0.344	*0.362
tense	0.128	0.058	0.066	0.058	0.081	0.096	0.089	0.104
lexical choice of noun	0.054	0.054	0.124	0.133	*0.189	*0.216	*0.250	*0.258
lexical choice of verb	0.098	0.098	0.087	0.138	*0.196	*0.232	*0.232	*0.241
pronoun	0.112	0.063	0.131	0.150	0.177	0.195	0.213	0.213
agreement	0.340	0.197	0.224	0.248	0.260	0.284	0.307	0.307
adjective	0.206	0.094	0.165	0.219	*0.413	*0.426	*0.426	*0.446
verb other	0.109	0.204	0.240	0.311	0.291	*0.340	0.308	0.340
adverb	0.333	0.286	0.286	0.302	0.333	0.349	0.349	0.349
conjunction	0.161	0.161	0.161	0.191	0.161	0.191	0.191	0.191
word order	0.048	0.093	0.093	0.091	0.091	0.091	0.091	0.136
noun other	0.200	0.154	0.286	0.286	*0.531	*0.490	*0.490	*0.490
auxiliary verb	0.083	0.160	0.160	0.083	0.083	0.160	0.160	0.160
other lexical choice	0.182	0.000	0.095	0.095	0.400	0.400	0.400	0.400
relative	0.154	0.285	0.154	0.154	0.154	0.154	0.154	0.154
interrogative	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Total	0.148	0.146	0.180	0.200	0.239	0.247	0.254	0.260

（評価はF値=適合率と再現率の調和平均。KJ=KJコーパス、Lang-8=Lang-8コーパスを訓練データに用い、フレーズベースの統計的機械翻訳を用いて誤り訂正システムを構築した）

	learner	correct
article	I like a chocolate very much.	I like _ chocolate very much.
lexical choice of noun	my cycle was injured, but i wasn't.	my bicycle was damaged, but i wasn't.

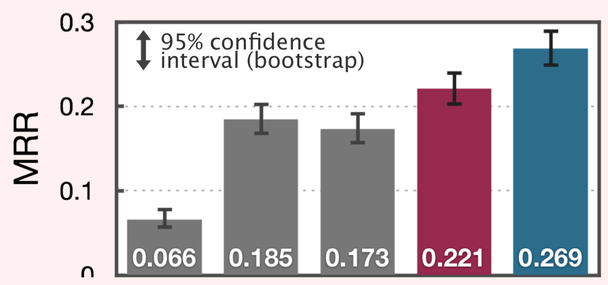
Table 5: Examples of system output for article and lexical choice of noun error

	learner	correct
noun number 1	I read various type books.	I read various types of books.
*noun number 2	There is a big snoopy dools in my room.	There is a big snoopy doll in my room.
tense 1	If I 'll live in saitama, I must have ...	If I _ live in saitama, I must have ...
*tense 2	The weather is very sunny, so we were ...	The weather was very sunny, so we were ...
agreement 1	Flowers is very beautiful.	Flowers are very beautiful.
*agreement 2	I think, reading comics are not "reading"	I think, reading comics is not "reading"

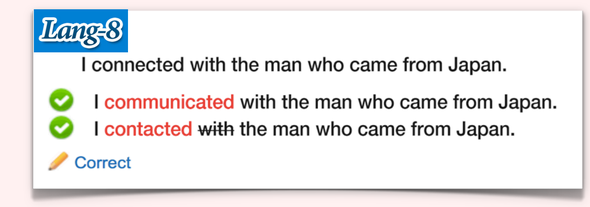
Table 6: Examples of system results for noun number, tense and agreement errors. Asterisks indicate that the SMT system using full Lang-8 Corpus failed to correct the errors.

誤りパターンを用いた動詞訂正実験

✓ 動詞の語彙選択誤り訂正では学習者の誤りパターンを考慮すると有意に性能が向上した



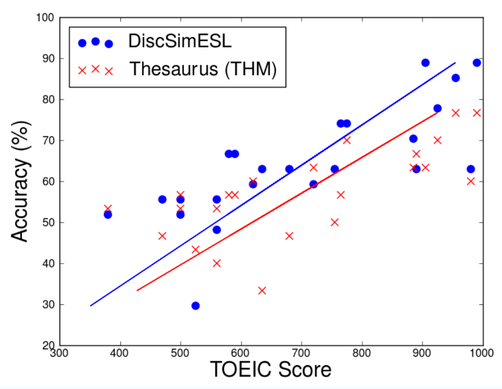
	WordNet	Roundtrip	WordNet+Roundtrip	Lang-8	Lang-8+EasyAdapt
カバー率	0.14	0.48	0.45	0.58	0.58



（評価はMRR=平均逆順位。動詞の語彙選択誤り訂正の候補を抽出する手法で比較。比較対象は類義語辞書を候補生成に用いたWordNet、対訳辞書を用いたRoundtrip、そしてそれらを組み合わせたWordNet+Roundtrip。誤り訂正の候補生成だけでなく、誤り訂正モデルを学習者コーパスを訓練データに用いて分野適応した提案手法がLang-8+EasyAdapt）

誤りパターンを用いた問題生成実験

✓ 動詞の穴埋め問題の選択肢の自動生成では学習者の誤りパターンを考慮すると有意に妥当な（非ネイティブの英語力と問題の正解率との相関が高い）問題が生成できるようになった



Each side, government and opposition, is _____ the other for the political crisis, and for the violence.

(a) blaming (b) accusing (c) BOTH

（評価は相関係数。3人の英語ネイティブに生成された問題を評価してもらい、2人以上の答えが一致した問題のみを用い、23人の日本人学生に生成された問題を解かせ、問題の正解率と申告されたTOEICのスコアの相関係数を調べた。比較手法はThesaurus=類義語辞書を用いて問題を生成する手法。提案手法DiscSimESLはLang-8から抽出された誤りパターンを用いて問題を生成する。）