ISS-P-126

CRF による英語時制誤り検出における有効な素性の検討

松井　悠太郎†　　新妻　弘崇†　　太田　学†
† 岡山大学大学院自然科学研究科

1. はじめに

普段英語を使用しない日本人が書いた英文には誤りが含まれることが多い．日本人英語学習者コーパスである Konan-JIEM Learner Corpus Edition (KJ コーパス)[1] のエラータグの統計では動詞に関する誤りが最も多く，その中でも時制に関する誤りは最多である．この結果をふまえ，本稿では英語の時制誤りに注目し，時制誤り検出に有効な特徴量について検討する．

2. 時制誤り検出の手法

本研究では，大域的な文脈を考慮した機械学習を用いる Tajiri らの手法[2]に注目する．具体的には，時制誤り検出を，文章中の動詞句に正しい時制ラベルを割り当てる系列ラベリング問題と捉えることで Conditional Random Field (CRF) を利用する．利用する素性の種類を表 1 にまとめる．これらは Stanford Parser 3.5.0[3]を用いて得た．nsubj, aux, prep, tmod, advmod は Stanford Parser の Typed Dependency をそのまま用いる．norm-p-tmod は p-tmod を「過去」「現在」「未来」「THIS」のいずれかに分類したものである．norm-p-tmod のラベルと対応するキーワードを表 2 に示す．なお，Tajiri らの手法では他に三種類の素性を用いているが，それらは本研究で用いた Stanford Parser の版では出力されなくなったため除いた．CRF で用いる素性テンプレートを表 3, 4 に示す．なお，表 4 で X' は対象動詞句の 1 つ前の動詞句の素性の種類 X を表す．

表 1： 動詞句の持つ素性の種類

素性の種類	内容
t-learn	学習者の書いた動詞の時制
head	動詞の原形
right	動詞の左の語
left	動詞の右の語
nsubj	動詞に係る主語
aux	動詞に係る助動詞
prep	動詞に係る前置詞
tmod	動詞に係る時間副詞（例:week）
p-tmod	動詞に係る時間副詞句（例:last week）
norm-p-tmod	正規化時間副詞
advmod	動詞に係るその他の副詞句（例:already）

表 2：norm-p-tmod

ラベル	キーワード（分類に利用）
過去	yesterday, last
現在	now
未来	tomorrow, next
THIS	today, this

表 3： 局所的素性テンプレート

<head> <head, t-learn> <head, L, R> <L> <L, head> <L, t-leran> <R> <R, head> <R, t-leran> <nsubj> <nsubj, t-learn> <aux> <aux, head> <aux, t-learn> <prep> <prep, t-learn> <tmod> <tmod, t-learn> <norm-p-tmod> <norm-p-tmod, t-learn> <advmod> <advmod, t-learn>

表 4： 大域的文脈素性テンプレート

<p-tmod'> <p-tmod', t-learn> <p-tmod', t-learn'> <p-tmod', t-learn', t-learn> <norm-p-tmod'> <norm-p-tmod', t-learn> <norm-p-tmod', t-learn'> <norm-p-tmod', t-learn', t-learn>

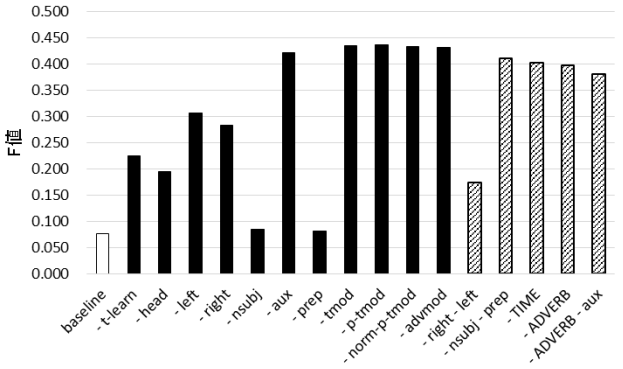

図 1： 特定の種類の素性を除いた時制誤り検出結果

3. 評価実験

実験では，CLC FCE Dataset[4]に収録されている 1,244 の英作文を用いる．このデータは 51,146 の動詞句を含み，そのうち 1,134 の動詞句が時制誤りである．ただし，動詞に限らず他の全ての誤りは修正している．本研究では工藤が作成した CRF++ 0.58[5]を用い，パラメータはデフォルト値を利用した．どの種類の素性が有効であるかを検討するため，特定の種類の素性を除いて時制誤り検出を行い，5 分割交差検定で F 値を求めた．その結果を図1にまとめる．図の baseline は表 1 で挙げた種類の素性を全て用いた結果，TIME は tmod，p-tmod，norm-p-tmod を除いた結果，ADVERB は TIME と advmod を除いた結果を表している．ただし，除く種類の素性を含むものは全て表 3，4から除く．図より，一種類の素性を除いた結果では，aux や tmod など動詞の時制決定に関係する素性を除くと F 値が大きく向上している．この原因として，学習者の書いた時制に誤りが多いことが予想される．複数の種類の素性を除いた結果では，一種類の場合と比べて F 値が下回ることが多い．しかし，－nsubj－prep のように素性を除く組み合わせによっては F 値が上がることを確認できた．

4. まとめ

本稿では，CRF による英語時制誤り検出に有効な素性について検討した．今後，時制の決定要因をさらに詳しく分析し，検出性能を向上させることを考えている．

参考文献

[1 KJ コーパス, http://www.gsk.or.jp/catalog/gsk2012-a/
[2] Tajiri, T., Komachi, M. and Matsumoto, Y., "Tense and Aspect Error Correction for ESL Learners Using Global Context," Proc. of ACL, pp. 198-202, 2012.
[3] Stanford Parser, http://nlp.stanford.edu/software/lex-parser.shtml
[4] CLC FCE Dataset, http://ilexir.co.uk/applications/clc-fce-dataset/
[5] CRF++, http://crfpp.googlecode.com/svn/trunk/doc/index.html