シンプルな規則による動詞誤り検出　および
周辺文脈を用いた分類器ベースの前置詞誤り検出システム

乙武　北斗* (ototake)
*福岡大学工学部
ototake@fukuoka-u.ac.jp

シンプルな規則による動詞誤り検出手法

概要

本システムは，シンプルな人手による規則を用いた
主語と動詞の一致に関する誤りを検出するシステムである.

利点

- トレーニングデータを必要としない

結果

	F	P	R
ドライラン	0.507	0.406	0.675
フォーマルラン	0.314	0.286	0.348

課題

- 主語の推定方法が不十分である点
  - Writing novels <v_agr>**is**</v_agr> very clear.
  - To play sports with many people <v_agr>**is**</v_agr> very interest for me.
- 疑問文に対応できていない点
  - what <v_agr>**do**</v_agr> you do your favorite things?
  - Why can the plane <v_agr>**do**</v_agr> it?

処理の流れ

She play tennis very well.

入力文

動詞句抽出

She **play** tennis very well.

主語と動詞を推定

She play tennis very well.

主語と動詞の対応をチェック

She **<v_agr>**play**</v_agr>** tennis very well.

動詞句抽出

- 入力文に対してトークン分割，品詞タグ付け，チャンキングを行う
- Apache OpenNLPライブラリを利用
- VP とラベル付けされた句を抽出する

主語の推定

1. 動詞句の前の句が名詞句であれば，その名詞句の主名詞が主語
   <subj>**She**</subj> <vp>**play**</vp> tennis very well.
2. 1.の推定が関係代名詞の場合，推定主語の一つ前の名詞が主語
   The <subj>**boy**</subj> who <vp>**play**</vp> tennis.
3. 1.の推定が there の場合は，動詞句の後ろの名詞句の主名詞が主語
   There <vp>**is**</vp> many <subj>**people**</subj>.

動詞の推定

1. 動詞句の最も先頭にある，POS がVB* の単語を人称変化動詞とする

主語と動詞の対応チェック

名詞／代名詞

I	1単
we	1複
you	2単複
he/she/it	3単
they	3複
NN/NNP	3単
NNS/NNPS	3複

人称・単複の対応チェック

am	1単
are/were	複
is	3単
was	単
VBP	NOT 3単
VBZ	3単

動詞

※
主語を含む名詞句に and を含む場合，
主語は複数形として扱う

周辺文脈を用いた分類器ベースの前置詞誤り検出手法

概要

本システムは，最大エントロピー分類器を用いた前置詞の置換・挿入誤りを検出するシステムである．
9種類の前置詞を検出対象としており，分類器が出力する前置詞の推定確率が閾値未満だった場合に，当該前置詞は誤りであると判断する．
対象とする前置詞は以下の9つである．

of, in, on, at, for, by, to, from, about

利点

- WordNetカテゴリや固有表現タイプを素性として用いることで，トレーニングデータに存在しない単語にもある程度対応できることを目指した．

結果

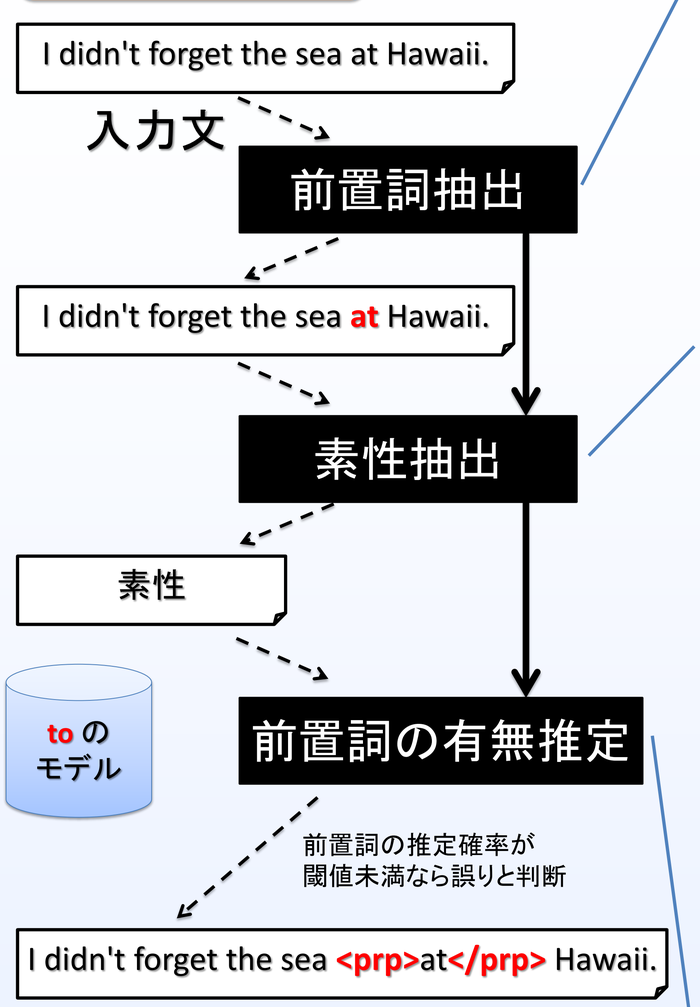
	F	P	R
ドライラン	0.287	0.232	0.378
フォーマルラン	0.305	0.343	0.275

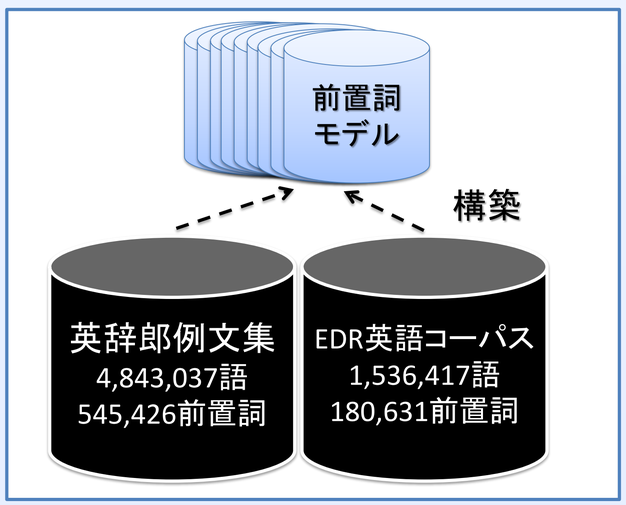
課題

- 前置詞の脱落誤りに対応できない点
  - ドライランのデータ中には，脱落誤りは全体の41%
  - 動詞の直後を対象に，前置詞の有無を推定させる方法を検討している
- False Positiveの率が高い点

前置詞	正例数	FP	FP/正例
in	270	124	0.459
to	248	72	0.290
of	156	25	0.160
for	112	45	0.402
at	65	33	0.508
by	61	46	0.754
on	49	28	0.571

処理の流れ





前置詞抽出

- 入力文に対してトークン分割，品詞タグ付け，チャンキングを行う
- 動詞誤りと同様，Apache OpenNLPライブラリを利用
- PP とラベル付けされた句から前置詞を抽出する

素性抽出

[参考] De Felice et al., "A classifier-based approach to preposition and determiner error correction in L2 English," Coling 2008

素性	例
1つ前の単語	sea
1つ前の単語の品詞	NN
1つ前の単語のWordNetカテゴリ	noun_object
1つ後ろの単語	Hawaii
1つ後ろの単語の品詞	NNP
1つ後ろの単語の固有表現タイプ	location
前後3単語の品詞	VB, DT, NN, NNP, .

- 固有表現タイプの推定には，解析でも用いている Apache OpenNLPを利用した．
- 推定可能なタイプは以下の7種類．
  - date, location, money, organization, percentage, person, time

前置詞の有無推定

[参考] N. Okazaki, "Classias: a collection of machine-learning algorithms for classification," http://www.chokkan.org/software/classias/

- 最大エントロピー分類器を利用
- 入力文に現れている前置詞に対応した二値分類器を用いて，推定確率を取得する．
- 確率が閾値未満の場合，その前置詞は誤りであると判断し，誤りタグを付与する．

閾値を変化させたドライランの結果

閾値	F
0.1	0.255
0.2	0.285
0.3	**0.287**
0.4	0.286
0.5	0.264