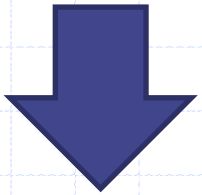
EDCW2012
前置詞誤り検出タスク成果報告

2012年9月3日
乙武 北斗 （ototake）
福岡大学工学部

タスク概要

- 日本人英語学習者による英文に対して前置詞誤り部分にタグ付けを行う

I didn’t forget the sea at Hawaii.



システム適用

I didn’t forget the sea **<prp>**at**</prp>** Hawaii.

誤りの種類

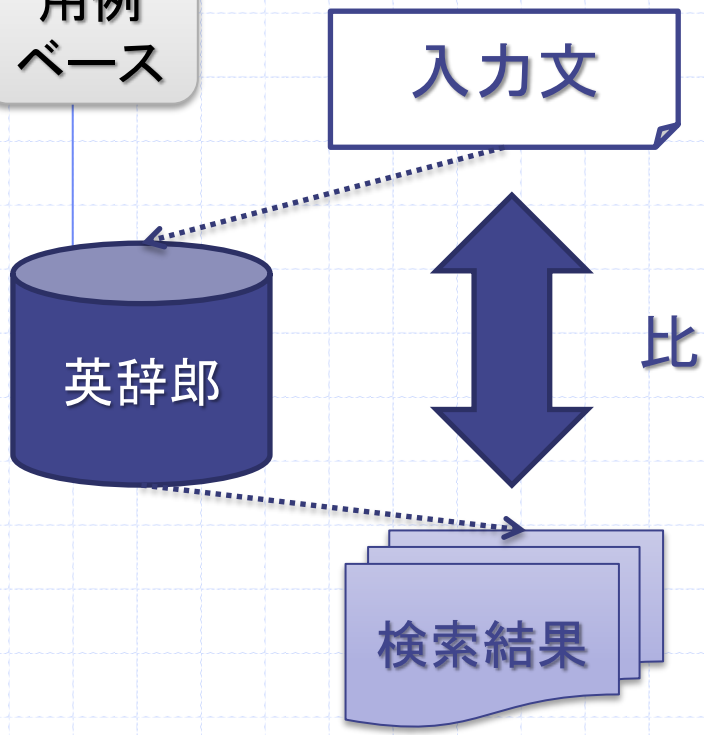
※数値はドライランデータ中の前置詞誤りタグの状況

- 脱落誤り（149 / 361）
  - 本来必要な前置詞が抜けてしまっている誤り
- 挿入誤り（99 / 361）
  - 本来不要な前置詞が挿入されている誤り
- 置換誤り（113 / 361）
  - 異なる前置詞による誤用

誤り検出対象とした

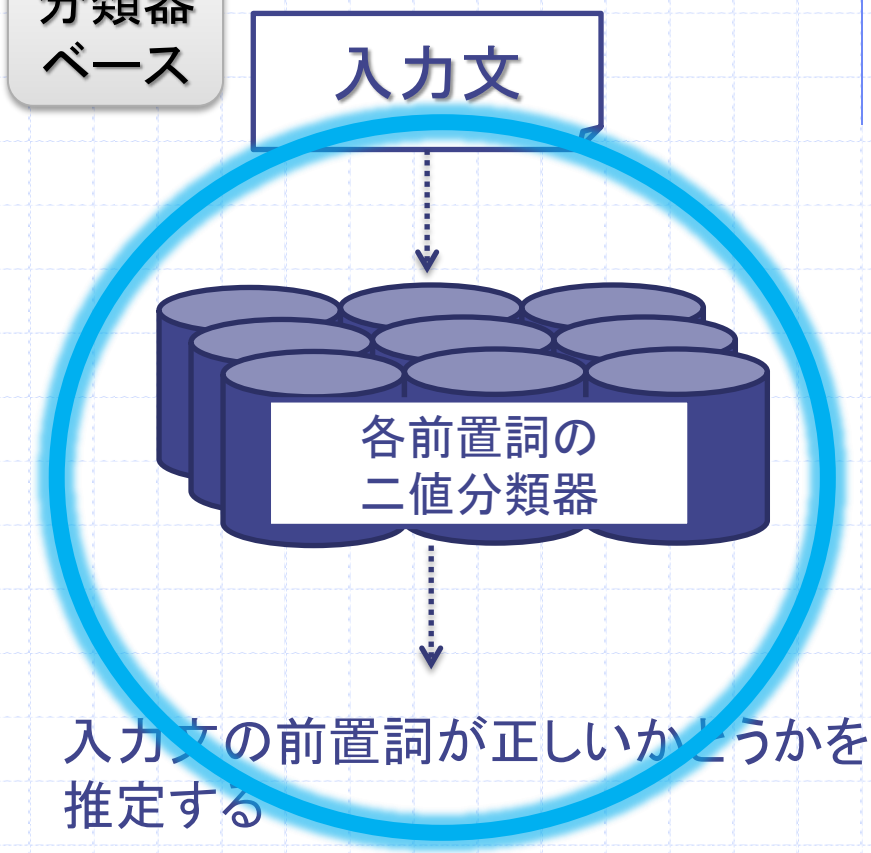
2つの異なるアプローチ

用例ベース



英辞郎から前置詞の用例を検索結果と比較して，正誤を推定

分類器ベース



入力文の前置詞が正しいかどうかを推定する

用例ベース手法 概要

I **left to** Aichi.

動詞句の直後に前置詞を伴うもの のみを誤り検出対象としている

対象前置詞
- of
- in
- on
- at
- for
- by
- to
- from
- about

"leave to"

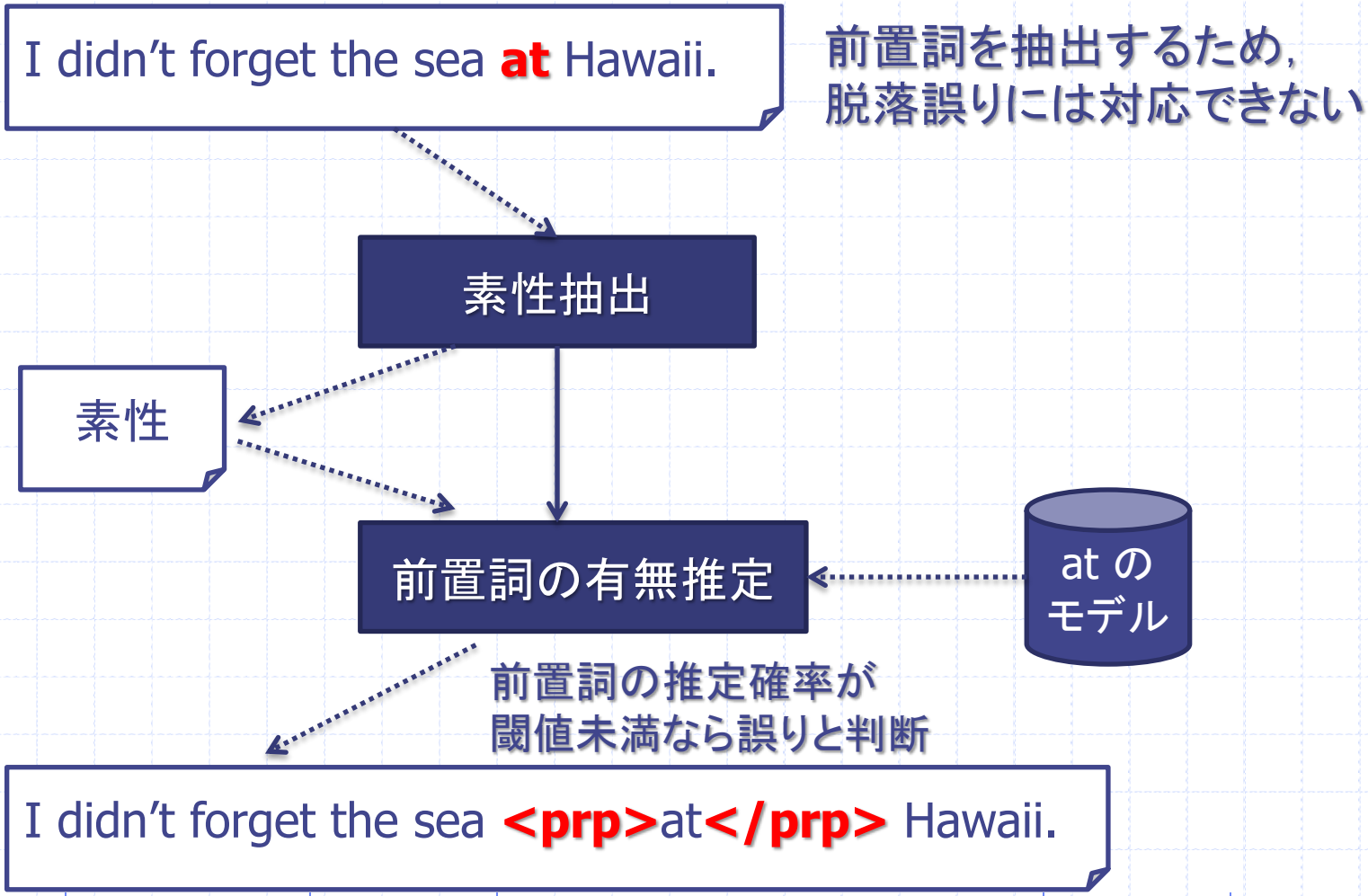
用例検索

英辞郎
182万項目

- **go shopping**
  買い物のため外出する
- **leave to go to the store**
  店に出掛ける
- **leave to revert to nature**
  自然のままにしておく、自然に
- **leave to rust**
  さびつくままにしておく
- **leave to rust and decay**
  さび朽ちるに任せる
- **leave to someone the final de**
  ～についての最終決定を（人）に任
- **leave to someone's option**
  任意とする

- 検索結果から
  leave to NP
  という表現がある項目を数える
- 全件のうち，上記表現のある割合が閾値未満の場合は誤りと判定

分類器ベース手法 概要


I didn’t forget the sea **at** Hawaii.
前置詞を抽出するため，
脱落誤りには対応できない
素性抽出
素性
前置詞の有無推定
at の
モデル
前置詞の推定確率が
閾値未満なら誤りと判断
I didn’t forget the sea <prp>at</prp> Hawaii.

分類器ベース手法
素性抽出

I didn't forget the sea **at** Hawaii.

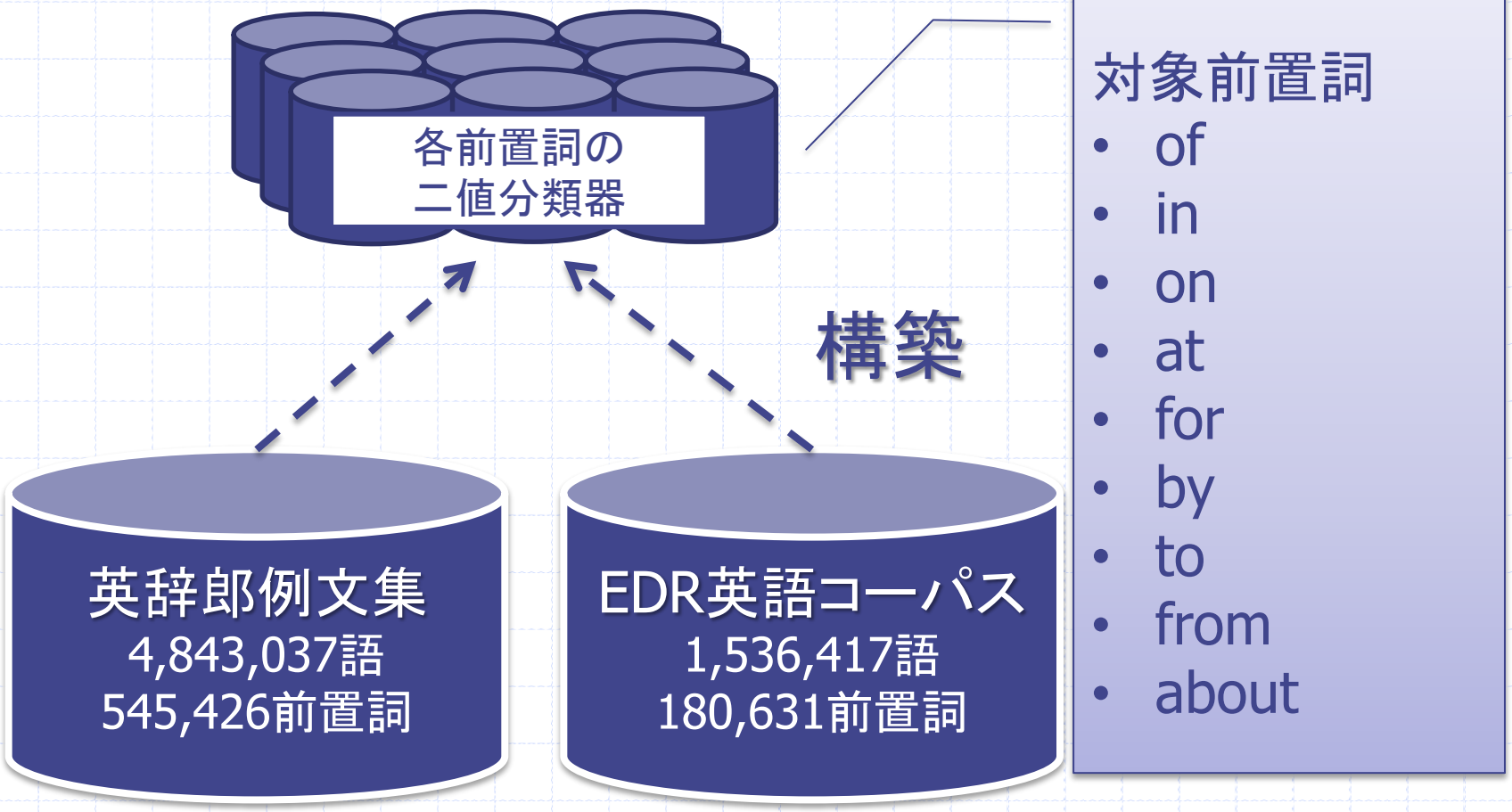
入力文から
前置詞を抽出

その周辺から
素性抽出

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, .

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

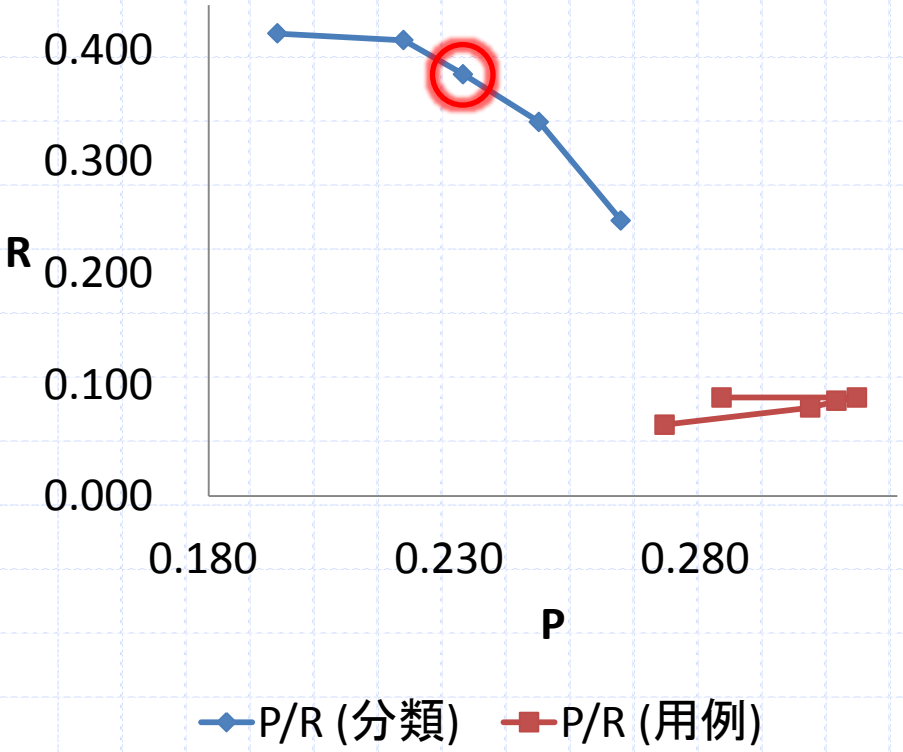
分類器ベース手法 モデルの構築



対象前置詞
- of
- in
- on
- at
- for
- by
- to
- from
- about

ドライラン評価
2つのアプローチの比較

閾値	F（分類）	F（用例）
0.1	0.255	0.104
0.2	0.285	0.126
0.3	**0.287**	0.134
0.4	0.286	**0.138**
0.5	0.264	0.135



2つのアプローチの比較詳細

前置詞	用例ベース	分類ベース
in	0.429 (3/7)	0.244 (40/164)
to	0.447 (21/47)	0.321 (34/106)
for	0.500 (3/6)	0.182 (10/55)
on	0.667 (2/3)	0.200 (7/35)

前置詞別Precision

前置詞	用例ベース	分類ベース
in	0.070 (3/43)	0.930 (40/43)
to	0.276 (21/76)	0.447 (34/76)
for	0.167 (3/18)	0.556 (10/18)
on	0.250 (2/8)	0.875 (7/8)

前置詞別Recall

- 用例ベースのみで検出できたもの
  - I'll continue to thank **<prp** crr="">to**</prp>** my mother.
  - My family visited **<prp** crr="">to**</prp>** the Shirahama ...
  - It took **<prp** crr="">for**</prp>** about eleven hours.

2つのアプローチの比較詳細

前置詞	用例ベース	分類ベース
in	0.429 (3/7)	0.244 (40/164)
about		0.125 (3/24)
to	0.447 (21/47)	0.321 (34/106)
from		0.111 (2/18)
at		0.353 (18/51)
for	0.500 (3/6)	0.182 (10/55)
by		0.137 (7/51)
of		0.107 (3/28)
on	0.667 (2/3)	0.200 (7/35)

前置詞別Precision

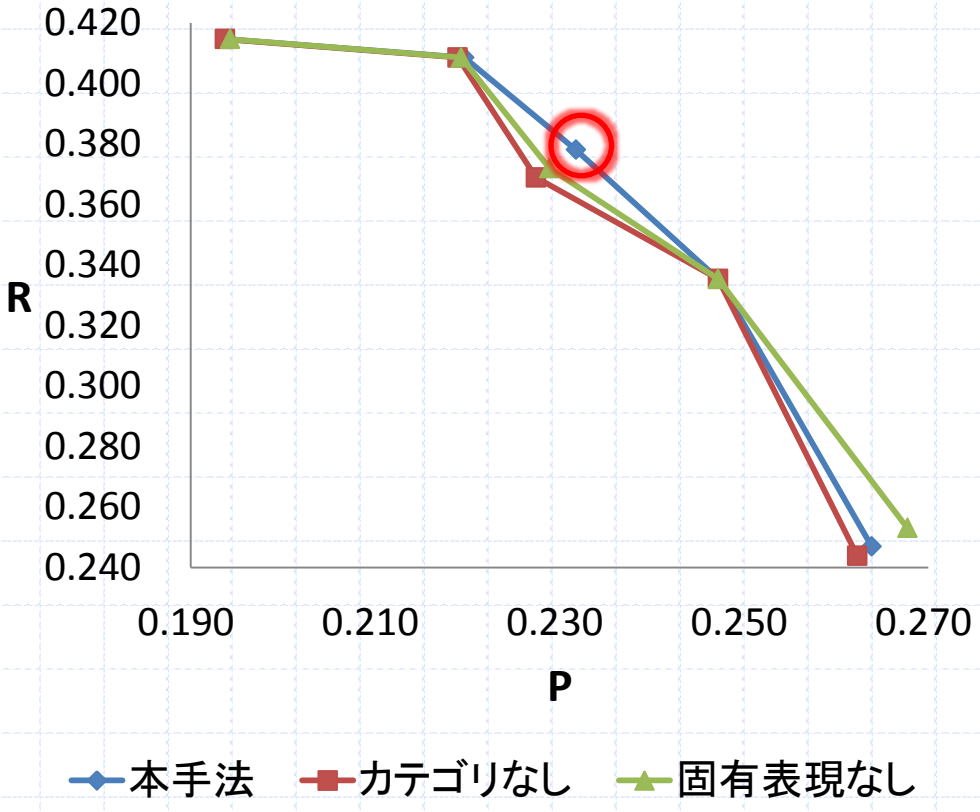
前置詞	用例ベース	分類ベース
in	0.070 (3/43)	0.930 (40/43)
about		1.000 (3/3)
to	0.276 (21/76)	0.447 (34/76)
from		0.667 (2/3)
at		0.900 (18/20)
for	0.167 (3/18)	0.556 (10/18)
by		0.875 (7/8)
of		0.500 (3/6)
on	0.250 (2/8)	0.875 (7/8)

前置詞別Recall

ドライラン評価
分類ベースの素性の違い

- F0
  - すべての素性を使用
- F1
  - WordNetカテゴリを除く
- F2
  - 固有表現タイプを除く

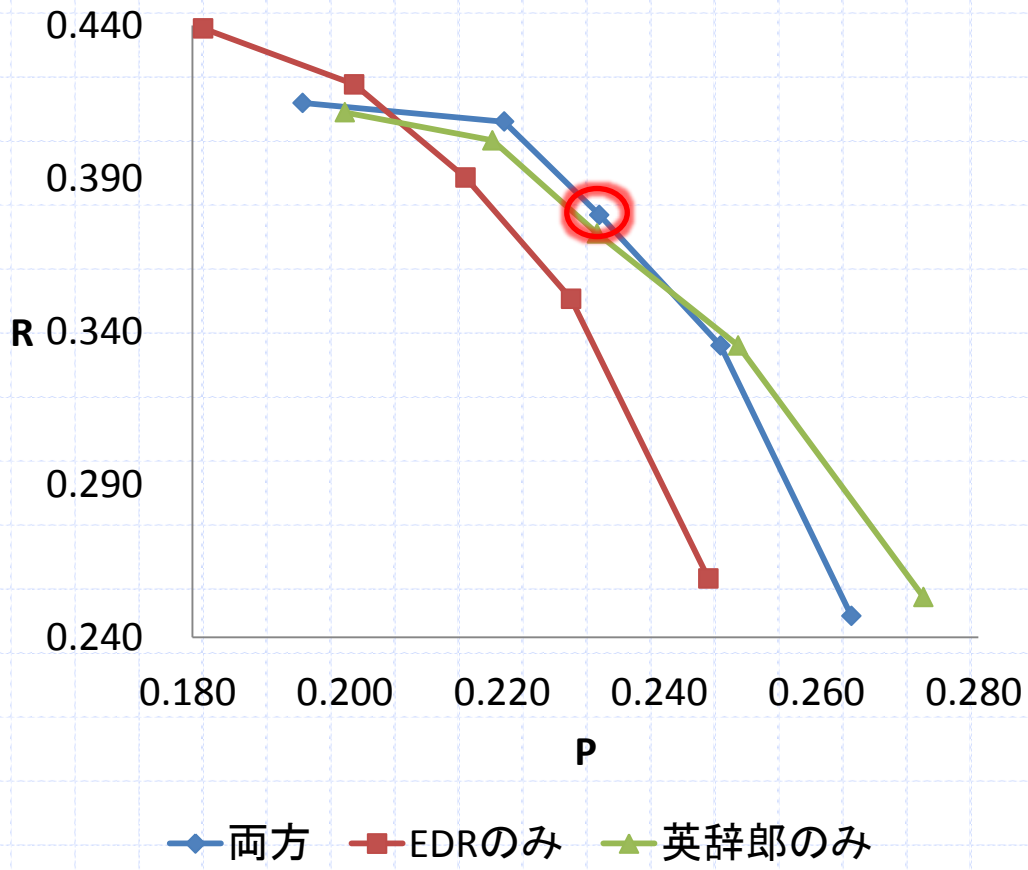
閾値	F0	F1	F2
0.1	0.255	0.253	0.260
0.2	0.285	0.285	0.285
0.3	**0.287**	0.281	0.283
0.4	0.286	**0.285**	**0.285**
0.5	0.264	0.264	0.265



ドライラン評価
トレーニングデータの違い

- F
  - EDRと英辞郎例文集
- Fe
  - EDRのみ
- Fj
  - 英辞郎例文集のみ

閾値	F	Fe	Fj
0.1	0.255	0.252	0.263
0.2	0.285	0.276	**0.286**
0.3	**0.287**	**0.277**	0.285
0.4	0.286	0.272	0.283
0.5	0.264	0.257	0.269



課題

- 前置詞の脱落誤りに対応する
  - 今回紹介した2つの方法ともに対応不可
  - 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中
- 脱落誤り（149 / 361）=0.413
  - 本来必要な前置詞が抜けてしまっている誤り

課題

- False Positiveの改善

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

ご清聴ありがとうございました.

ポスターでは，動詞と前置詞誤りの
両方について展示いたします.