NAIST Error Detection System at EDCW 2012

水本智也, 林部祐太, 坂口慶祐
奈良先端科学技術大学院大学
(tomoya-m, yuta-h, keisuke-sa)@is.naist.jp

前置詞誤りの検出

❖ 提案手法

12 種類の前置詞: of, in, for, to, by, with, at, on, from, as, about, sinceの置換誤りと挿入誤りの検出を行う.

以下の素性を，前置詞が出現した箇所に対して，抽出し，学習事例を作成する. ラベルは訂正後の前置詞である. これを，最大エントロピー法を用いた多クラス分類器の学習データとして用いる.

誤り検出時には，前置詞が出現した箇所に対して抽出し，事例を作成する. そして先に学習した分類器で，最尤前置詞を予測する. それが作文者の用いた前置詞と異なっていれば，誤りと判定する.

❖ 素性の一覧

- 表層素性
  - 周辺 2 単語の表層形・品詞・WordNet の意味クラス（全 40 種類）
  - 先行する動詞句・名詞句の主辞
  - 後続する動詞句・名詞句の主辞
- 構文素性
  - 前置詞の主辞・補語の表層形・品詞
  - 前置詞と主辞・補語の関係
  - 前置詞の構文木上での親と親の親のノード名
  - 前置詞の構文木上での親の子のノード名
- 頻度・意味素性
  - 着目している前置詞を prep に置換し，左 i 単語の列L と，右 j 単語の列 R を用い (i + j ∈ { 3,4,5 })，単語列 LprepR を Web N-gram コーパスで検索したときの頻度f_prep
  - f_prepと，訂正前置詞候補それぞれのfの商

全ての誤りの検出

❖ 統計的機械翻訳を用いた誤り訂正手法

1. 学習者コーパス（学習者の書いた文とその添削文）から誤り訂正モデルの学習を行なう.
2. ネイティブの書いた英文コーパスから言語モデルを学習する（3-gramを使用）.
2. 誤り訂正モデルと言語モデルを使って，学習者の文に対して誤り訂正を行なう.
3. 元の学習者の文とシステムの出力を比べ（動的計画法を使用），システムが訂正を行なった箇所を誤りとして検出する.
4. 前置詞誤り訂正システムの出力結果と動詞の一致誤り訂正システムの出力をマージする.

❖ トレーニングデータ

・言語学習SNS Lang-8*からクローリングして獲得したデータを使用
・日本人英語学習者の作文とその添削文の509,116文対
・文の構造が大きく変わるような添削の場合，アライメントを失敗しやすいため，挿入数，削除数5以下（動的計画法で計算）のもののみを抽出し最終的に391,699文対を使用.

*http://lang-8.com

動詞の一致誤りの検出

❖ 動詞の一致とは?

文中において主語の人称，数，格，時制とそれに呼応する動詞の形が一致しなければならない，という文法規則

❖ 提案手法

1. 入力文を Stanford Parser で解析し，依存関係を取得する.
2. 1. の結果から,動詞に対する主語を取得． 追加規則（表1）が適応できる場合は追加規則に従う.
3. 主語と動詞の人称，数が一致しているか確認する.
4. 一致していない場合は誤りとして検出する.

表1. 追加規則

追加処理する項目	追加処理の詳細	例文
受動態	受動態の場合は過去分詞形となる動詞ではなく，その動詞に係る受動態助動詞 (auxpass) を一致の対象とする.	I *were / was impressed ...
関係詞	関係詞 (WDT) が主語と解析された場合は，その先行詞 (関係詞に先行する名詞) を一致対象の主語とする.	... power that *make / makes people happy.
並列句構造	主語が並列句構造であるかどうかは，主語の後ろに *and* もしくは ',' が来ているかどうかで判定する． 並列句構造であると判定された場合，主語は複数形であるとみなす.	There *is / are a farm and many caws.
動名詞	主語が *-ing* で終わる場合，元々の品詞に関わらず，単数名詞とみなす.	*Reading* books is very good ...
数量を表す名詞	主語が *lot, many, more, most* と解析された場合は，元々の品詞に関わらず複数名詞とみなす.	There *was / were a lot of people.
主語の誤り	主語が単数形で，前 3 単語以内に数詞がある場合は主語を複数名詞とみなす.	There were two little *garden /gardens ...

用いた主な外部ツール・データ

Stanford Parser	構文解析器
Jpype	JavaのPythonラッパ
NLTK	言語処理ツール群
Maxent	最大エントロピー分類器
SSGNC	Ngram検索ツール
SSGNC-python	SSGNCのpythonラッパ
Web 1T 5-gram	N-gramデータ
Moses	機械翻訳デコーダ
SRILM	言語モデル構築
GIZA++	単語アライメント

結果

	F-measure	Precision	Recall
動詞の一致	0.689	0.553	0.913
前置詞	0.337	0.479	0.260
全て	0.361	0.484	0.288

参考

・水本智也, 林部祐太, 坂口慶祐, 小町守, 松本裕治. 英作文誤り訂正における複数の手法の利用に関する考察. 情報処理学会第208回自然言語処理研究会, Vol.2012-NL-208, September 2012.