システムディスクリプション

|  |
| --- |
| 記入日：2012年8月9日  対象トラック：　前置詞トラック  システム名：　ME分類ベースシステム  チーム名：　ototake  性能：　ドライラン　 R=0.378　 P=0.232　　F=0.287  　　　　フォーマルラン　R=0.275　 P=0.343　 F=0.305 |

1. 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0.1未満だった場合に，当該前置詞は誤りであると判断する．

1. 実行方法

実行環境・使用言語：Windows .NET Framework 4.0 または Linux Mono 2.10，C#

実行方法：（詳しくはREADME.txtを参照）

1. Classiasがインストールされていて，パスが通っていることを確認する
2. OpenNLPモデルファイル，WordNetデータベースファイルを適当なディレクトリにに置く（以下，~/resとする）
3. 前置詞推定モデルファイルを適当なディレクトリに置く（以下，~/modelsとする）
4. KJ Corpusデータを適当なディレクトリに置く（以下，~/dataとする）
5. 誤り検出  
   Console.exe d --target prp --kjdir ~/data/corpus\_data --formal --res ~/res --out output\_dir --model ~/models
6. 使用データ，ツール，辞書

ツール名：IKVM.NET

　　URL：http://www.ikvm.net/

種別：仮想マシン（フリーソフトウェア）

用途：JavaプログラムであるApache OpenNLPの.NETアセンブリへの変換

ツール名：Apache OpenNLP

　　URL：http://opennlp.apache.org/

種別：解析器（フリーソフトウェア）

用途：トークン分割，POSタグ付け，Chunking，固有表現抽出

ツール名：SGMLReader

URL：https://github.com/MindTouch/SGMLReader

種別：SGML解析ライブラリ（フリーソフトウェア）

用途：KJ CorpusのXML解析

ツール名：DotNetZip

URL：http://dotnetzip.codeplex.com/

種別：ZIPライブラリ（フリーソフトウェア）

用途：結果出力用のZIP圧縮

ツール名：ConsoleLib

URL：https://bitbucket.org/shokai/consolelibnet

種別：コマンドパラメータ解析ライブラリ（フリーソフトウェア）

用途：コマンドラインオプション解析

ツール名：Classias

URL：http://www.chokkan.org/software/classias/index.html.ja

種別：機械学習実装（フリーソフトウェア）

用途：前置詞推定モデル構築および分類

データ名：EDR電子化辞書 英語コーパスVer. 2.0

　　URL：http://www2.nict.go.jp/out-promotion/techtransfer/EDR/J\_index.html

種別：英語コーパス（有償）

用途： 前置詞推定モデル作成

データ名：英辞郎 Ver.128 例辞郎

　　URL：http://www.eijiro.jp/

種別：英語例文集（有償）

用途：前置詞推定モデル作成

辞書名：WordNet

　　URL：http://wordnet.princeton.edu/wordnet/download/current-version/

　　用途：Lemmatize，意味カテゴリ推定

1. システムの詳細（システムの詳細を記入ください）

処理の流れは以下の通り．

対象とする前置詞は

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

の9つ．

対象とする誤りは，置換誤りと挿入誤り．

トレーニングデータより前置詞と後述する素性を抽出し，誤り対象前置詞各々の二値分類モデルを最大エントロピー分類器として構築する．

誤り検出の際は，前置詞の推定確率が0.1未満の場合に誤りと判定し，誤りタグを付与する．

用いた素性は以下の通り．

* 前置詞の1つ前の単語，POS，WordNetカテゴリ
* 前置詞の1つ後ろの単語，POS，固有表現タイプ
* 前置詞の前後3単語のPOS

固有表現タイプの推定にはOpenNLPを利用した．推定可能なタイプは以下の通り．

date, location, money, organization, percentage, person, time

1. セールスポイント（セールスポイントを記入ください）

* 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．