システムディスクリプション

|  |
| --- |
| 記入日：2012年8月12日  対象トラック：　オープントラック  システム名：　検索エンジンの検索結果数を用いた誤り検出システム  チーム名：　okayamaU  性能：　ドライラン　R=0.248 P=0.183　F=0.211  　　　　フォーマルラン　R=　P=　F= |

1. 概要

本システムは，検索エンジンの検索結果数に基づき，英文中の誤りを検出するシステムである．誤りを含む英文やフレーズを検索すると，その検索結果数は小さくなることが予想できる．本システムでは，その性質を利用して英文中の誤りを検出する．まず，検討する英文から単語n-gramを生成し，それらの検索結果数を求める．求めた検索結果数を素性として，SVM (Support Vector Machine) による誤り検出を行う．

1. 実行方法

実行環境・使用言語：Windows7・Java

実行方法：

1. 前準備

stanford-parser.jar, stanford-parser-2012-05-22-models.jarをclasspathに追加

Yahoo!Web検索APIキーを取得し，Search\Search.javaに記述

TinySVMをダウンロードし，Windowsのパスを通す

データを適当なディレクトリに入れる（以下dataとする）

1. コンパイル

javac detect/\*.java search/\*.java

1. 誤り検出

java detect.Main –i data

1. 使用データ，ツール，辞書

ツール名：Stanford Parser (Version 2.0.2)

　　URL：http://nlp.stanford.edu/software/lex-parser.shtml

種別：構文解析器

用途：トークナイゼーション

ツール名：TinySVM 0.09

　　URL：http://chasen.org/~taku/software/TinySVM/

種別：SVM 学習パッケージ

用途：誤り検出

ツール名：Yahoo!検索Web API（アップグレード版）

　　URL：http://developer.yahoo.co.jp/

種別：Web API

用途：検索結果数の取得

データ名：アメリカの子どものように英会話を覚える本

種別：書籍

用途：学習データ

1. システムの詳細

　システムの簡単な処理の流れを説明する．まず，英文から単語n-gramの検索クエリを生成して，フレーズ検索により，検索結果数を得る．その検索結果数に基づいて素性を定め，SVMにより英文中の誤りを検出する．本システムでは，英文中の誤りを検出するために，SVMを二回使用する．一回目で検討する英文が誤りを含んでいるか判別し（誤り文検出），二回目で英文中の誤っている語を検出する（誤り単語検出）．ここで，誤り文検出で英文が誤りを含んでいると判定しても，誤り単語検出で誤った単語を検出できなかった場合は，誤りはないと判断する．

　まず，誤り文検出の方法について説明する．英文に含まれる語数はわからないので，英文の先頭から順に単語n-gramを生成し，そのいずれかを誤りと判定した場合，英文が誤っているとする．具体的には，検討したい英文の先頭から順に4-gramを生成する．ただし，文末のみtrigramとする．次に，それらの単語列が誤りを含むかどうかSVMを用いて判定する．いずれかの4-gramまたはtrigramが誤りを含むと判定された場合，その文は誤りを含むと判断する．例えば，“I had a dream last night.”という英文からは，“I had a dream”, “had a dream last”, “a dream last night”,“dream last night”の四つの単語列を生成し，これらの単語列が誤っているかどうか判定する．

　誤り文検出に用いた素性を表 1にまとめる．表 1は検討する英文のi番目の単語を先頭に持つ4-gramやtrigramなどに基づいて算出する素性の一覧で，これらがi番目の単語から始まる単語列の特徴ベクトルを構成する．ここで，はi番目の単語を先頭に持つn-gramの検索結果数である．例えば，特徴ベクトルの第1成分はtrigramの検索結果数を，それを構成する三つのunigramの検索結果数の積で割った数の対数をとったものである．“I had a dream”について考えると，“I had a”の検索結果数を“I”，“had”，“a”の検索結果数の積で割った数の対数である．同様に，第2成分は“I had a dream”の検索結果数を“I”と“had a dream”の検索結果数の積で割った数の対数である．このような素性を英文から生成した4-gram, trigramに対して求める．なお，trigramに対しては特徴ベクトルの第2成分と第3成分は計算できないので値を0とする．なお，学習データには「アメリカの子供のように英会話を覚える本」に収録されている誤り文153文と，その正解例153文の計306文を用いた．

　次に，誤り単語検出について説明する．誤っている単語の検出に用いた素性を表 2にまとめる．表 2は検討したい英文のi番目の単語の特徴ベクトルを構成する素性の一覧を示している．ここでのも表 1のそれと同様の意味をもつ．また，文頭や文末付近の単語に対して特徴ベクトルを求める際，計算できない素性もある．具体的には，素性の計算中に出現するいずれかのにおいて，が負の数になる場合と，が文中の単語を超える場合である．そのような素性については値を0とする．表 2の特徴ベクトルを英文中の各単語に対して求め，SVMを用いて誤っている単語を検出する．また，誤り単語検出に用いた学習データは，誤り文検出の学習に用いた英文と同じものである．ただし，正解例は使わず，誤りを含む英文153文のみを用いている．

1. セールスポイント

　学習に単語などの情報を使わず，検索結果数のみを用いている．そのため，学習データが少なくても誤り検出を実行できる

表 1　誤り文検出に用いる素性

|  |  |
| --- | --- |
| Feature | Value |
| 1 |  |
| 2 |  |
| 3 |  |

表 2　誤り単語検出に用いる素性

|  |  |
| --- | --- |
| Feature | Value |
| 1 |  |
| 2 |  |
| 3 |  |
| 4 |  |
| 5 |  |
| 6 |  |
| 7 |  |
| 8 |  |
| 9 |  |
| 10 |  |
| 11 |  |
| 12 |  |
| 13 |  |
| 14 |  |
| 15 |  |
| 16 |  |
| 17 |  |
| 18 |  |
| 19 |  |
| 20 |  |
| 21 |  |
| 22 |  |
| 23 |  |
| 24 |  |
| 25 |  |
| 26 |  |
| 27 |  |
| 28 |  |
| 29 |  |
| 30 |  |
| 31 |  |
| 32 |  |
| 33 |  |
| 34 |  |
| 35 |  |
| 36 |  |
| 37 |  |