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：　動詞トラック  システム名：　品詞タグと検索結果数に基づく動詞誤り検出システム  チーム名：　okayamaU  性能：　ドライラン　R=0.614 P=0.429　F=0.505  　　　　フォーマルラン　R=　P=　F= |

1. 概要

本システムは，品詞タグとチャンク情報に基づき，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．また，品詞タグとチャンク情報から判断することが難しい英文に対しては，検索エンジンから得られる検索結果数を用いて，主語と動詞の一致に関する誤りを検出する．

1. 実行方法

実行環境・使用言語：Windows 7，Java

実行方法：

1. 前準備

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

データを適当なディレクトリに入れる（以下dataとする）

Yahoo! Web検索APIキーを取得し，search\Search.javaを編集

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：品詞タグ，チャンク情報の取得

データ名：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

1. システムの詳細
   1. システムの処理の流れ

　システムの簡単な処理の流れは以下のようになる．

1. 品詞タグ，チャンク情報から主語-動詞の対応関係の取得
2. 動詞の活用形の判断
3. 主語の単複・人称の判断
4. (2)，(3)が一致していなければ動詞にエラータグを付与

　なお，(3)で主語の単複・人称を判断することが難しい場合がある．例えば，主語に動名詞が含まれている場合，形容詞的用法か名詞的用法であるか判断しなければならない．このような場合，検索エンジンから得られる検索結果数を用いて誤りを検出する．

　検索エンジンを用いた誤り検出の流れは以下のようになる．

1. 主語と対応する動詞からなる検索クエリの生成

1. (1)の動詞の活用形を変えた検索クエリの生成
2. (1)，(2)それぞれの検索クエリでフレーズ検索し，検索結果数を取得
3. (1)の検索結果数より(2)の検索結果数が大きい場合，エラータグを付与
   1. 主語-動詞の対応関係の取得

まず，誤り検出の対象となる動詞とそれに対応する主語を探す．本システムでは，検討したい英文中に以下のパターンでチャンクが出現した場合，VPに検出対象の動詞が含まれている可能性があると判断する．

1. NP + VP (名詞句 + 動詞句)
2. NP + ADVP + VP (名詞句 + 副詞句 + 動詞句)
3. WHNP + VP (関係代名詞句 + 動詞句)

ここで，各パターンに含まれているNPが主語とは限らないが，その判断は後の処理で行う．

* 1. 動詞の活用形の判断

　次に，品詞タグに基づき動詞の活用形が，三人称単数現在形であるかそれ以外であるかを調べる．対象となる動詞は4.2で見つけたVPの先頭の単語である．VBZであった場合は三人称単数現在形であり，VB, VBPであった場合はそれ以外である．なお，4.2のVPチャンクの先頭の単語の品詞タグがVB, VBP, VBZ以外の場合は，このチャンクに対する誤り検出を行わない．

　ただし，VPの先頭の単語の品詞タグがRB (副詞) の場合は，2単語目の単語の品詞タグより動詞の活用形を判断する．また，VPの先頭の単語が “was”, “were” の場合は，“was” を三人称単数の主語に対応する動詞，“were” をそれ以外と判断する．

* 1. 主語の単複・人称の判断

4.2で得たNPからは主語の数・人称が正しく得られない可能性がある．例えば，NP + PP + NP + VPのようなパターンでチャンクが並んでいる場合が考えられる．ここで，PPは前置詞句を示す．このようなパターンは，VPの直前のNPが前置詞によって最初のNPを修飾している場合にみられる．そのため，PP + NPというチャンクのパターンが出現したら，その前にあるNPを主語とみなす．また，4.2の③のパターンに適合した場合，WHNPの直前のNPを主語とみなす．その際，WHNPの直前のNPについても，NP + PP + NPというパターンでチャンクが出現した場合は，PPの前を主語と見なす．また，There is構文の場合も，4.2で得たNPからは主語の単複・人称が判断できない．そのため，NPの最後の単語が“there”場合，VPの次のチャンクを調べ，そのチャンクがNPであった場合，そのNPチャンクを主語とみなす．

次に，主語が三人称単数かそれ以外であるかを調べる．まず，主語とみなしたNPの前後のチャンクを調べ，カンマで区切られたNPが2つ以上ある場合，名詞が列挙されているとみなし，主語は複数と判断する．また，NPに“and”が含まれており，その前後に名詞がある場合も，複数と判断する．また，主語とみなしたNPの前後のチャンクに“a lot of”という表現が含まれている場合も，複数とみなす．これらの条件に当てはまらない場合は，NPの最後の単語の品詞タグより判断する．すなわち，主語とみなしたNPの最後の単語の品詞タグがNN，NNPの場合は単数，NNS，NNPSの場合は複数とみなす．

また，NPがPRP (代名詞)，DT (限定詞) からなる場合がある．このような場合は単語を見て判断する．品詞タグがPRPである場合は，“he”，“she”，“it”，“one”を三人称単数とし，その他の単語は三人称単数でないと判断する．ただし，4.3で“was”を三人称単数の主語に対応する動詞としたので，“I”については例外処理をしている．また，DTの場合は，“these”，“those”を複数として，それ以外の場合は扱わない．

最後に，ここで，求めた主語の数・人称と4.3で求めた動詞の活用形を比較し，対応が正しくない場合，動詞に誤りタグを付与する．

* 1. 検索エンジンを用いた誤り検出

　4.4で述べた方法では，主語の数・人称を判断することが難しい場合がある．例えば，動名詞では形容詞的用法か名詞的用法か判断しにくい場合がある．そのような場合，本システムでは，検索エンジンを用い検索結果数を比較することで，誤りを検出する．

具体的には，4.4で主語と判断したNPに“ing”で終わる単語が含まれている場合と，主語としたNPの直前のチャンクがVPであり，なおかつ，品詞タグがVBGの単語1語からなる場合である．例えば，主語が“Reading books”だった場合，その品詞タグとチャンクは [NP Reading/NN books/NNS ]か，[VP Reading/VBG ] [NP books/NNS ]となることが多い．前者がNPに“ing”で終わる単語が含まれる場合であり，後者がNPの直前にVBGのみからなるVPがある場合である．これらのパターンで品詞タグとチャンクが出現した場合，検索結果数を用いて検出する．

まず，4.4で主語と判断したNPと4.2で求めたそれに対応するVPに含まれる単語から検索クエリを生成する．この際，NPの直前のVBGのみからなるVPに含まれる単語も検索クエリに含める．

次に，動詞の活用形のみを変化させた検索クエリを生成する．ここで活用形を変える動詞は4.3で求めた動詞と同じ動詞である．この動詞の品詞タグがVBZなら原形に，VB, VBPの場合は三人称単数現在形に変化させる．ただし，動詞がbe動詞であった場合は，原形の代わりに“are”を使う．また，be動詞の場合は，過去形にも対応する．すなわち，検討する英文の動詞が“was”の場合は“were”に，“were”の場合は“was”に変化させる．例えば，“Reading books is　～”という英文では，本システムは“Reading books is”と“Reading books are”という検索クエリを作成する．

こうして二つの検索クエリを生成し，フレーズ検索により検索結果数を取得する．得られた検索結果数を比較し，動詞の活用形を変えた検索クエリによる検索結果数のほうが大きい場合，エラータグを付与する．

1. セールスポイント

主に，チャンク情報と品詞タグを用いているので，特別に学習などをする必要がない．また，チャンク情報と品詞タグでは誤り検出が困難な場合に対応できるように，検索エンジンから得られる検索結果数を用いている．よって，誤りを含む英文を対象とするのでチャンク情報や品詞タグが誤っていることもあるが，検索エンジンを用いる部分に関しては，品詞タグなどの誤りをある程度許容できる．

また，検索エンジンを用いる条件や検索クエリ生成法を見直すことにより，誤り検出がより困難な英文にも対応することが可能であると考えている．