５６０２

web 検索サイトと n-gram を使った単語間の類似度計算の最適化

An optimization of calculating word similarity using web search engine and n-gram

CS29　林孝明
指導教員　大島真樹

1. はじめに

現在におけるチャットとは，データベースをつくりたくさんの単語を入力されている．これにより，実際に話しているような雰囲気を出すことが可能になったが，これらのデータベースは人間が手打ち，つまり手動で作っているためデータベースの中に入っているものはある程度の限定されたものしか入っていない．

そこで今回の行う研究としては，データベースの自動化をするために web 検索サイトを利用して検索結果を抽出し，n-gram を使用して単語間の類似度計算をすることにより人工無能等で使われているデータベースの内容更新を自動化する方法の提案をする．使用する web ページによってどのような違いが生まれてくるかを考察するまた，類似度計算を二つ以上の提案をしてどの提案が有効であるのかを考察する．

2. 研究方法

web サイトの検索動作をコンピュータプログラムで自動制御できるようにし，検索する単語を検索ボックスの中へ入れ，検索ボタンを押すことが出来るようにしておく．検索結果が表示されたら，プログラムによって web ページの解析を行い n-gram により統計情報を出す．n-gram により出された統計情報をもとにして解析し単語間の類似度を自動計算させる．類似度計算によって出た結果が正しいのかを判断するために類義語辞書により自動検証する．これらの結果が一通り終了したら類似度計算及び自動検証の結果についてデータベースに記録しておき再利用できるようにしていく．

3. 処理の流れ

まず入力をした文章より使われている単語を使い検索をかける．次に，検索エンジンのサイトへアクセスをかけ，テキストボックスに単語を挿入し，検索ボタンを押し検索させる．検索結果の表示がされたところで，ほかの単語を抽出して再検索をかけていく．この検索結果を用いて，n グラムや単語辞書を使って統計情報の抽出を行う．抽出した統計情報をもとに類似度計算を行ってそれぞれの結果より特徴や違いを出しデータベースに記録する．データベースは以下のようになっている．

表 1　データベース内の表の仕様

ID	単語1	単語 2	類似度 1	類似度 2
1	うどん	麺	70.3	62.4
2	うどん	油揚げ	63.3	50.8

ここで表記されているのはあくまで例であって結果ではない．

4. 結　果

今回 web 検索で使用した google と yahoo において検索を行った結果，yahoo で検索したときの結果が google で検索したときよりも類似度が高くなった．理由としては，web 検索をして類似度を計算し上位 10 個を比べたとき，始めの言葉から yahoo の検索結果が上回っていることが多かったからである．

ただし，欠点として変換している文字コードの範囲からはみ出すもの特に漢字が変換されず「？」となってしまう．

5. 結　論

本研究で web 自動検索を用いて n-gram で単語間の類似度計算により web サイトを別々使ってみることによって，類似度の変化が見ることが出来た．

類似度計算の方法については，別の方法を試せていないので行っていない．

6. 今後の発展

今回の提案した類似度計算の有効な方法が他にあると思われる．

また，検索サイトを利用して文字を抽出しているが指定している範囲以外の漢字をどのように反映させるかとテキスト抽出のされない場合の改良を検討していきたい．

参考文献

[1] 連想検索エンジン reflexa, http://labs.preferred.jp/reflexa/
[2] ビジネス用語辞典,Wisdom http://www.blwisdom.com/word/key/000876.html
[3] yahooJAPAN, http://www.yahoo.co.jp/
[4] google, http://www.google.co.jp/