背景と目的

文章間の類似性には様々な解釈があり，IR の分野などで様々な類似性評価の確立が重要となっている．しかしその中で他人の文章の真似であるかどうかに着目した研究は少ない．本研究では，他人の文章を真似して作成された文章を発見するため，部分的な類似性や類義語を考慮した類似性の評価法を提案する．本手法により，同一テーマで記述された文章と真似して書かれた文章とを明確に区別することができることを示す．

真似の指標計算法

本研究では 2 つの文章 X，Y 間の類似性を以下の手順により評価する．

1. 2 つの文章 X，Y を文に分割する．
2. 各文に対し形態素解析を行い，各文に含まれる名詞と動詞を取り出す．
3. 文章 X を構成する文と文章 Y を構成する文の全ての組み合わせに対し文間の類似性(距離）を計算する．各文は類義語を考慮した単語の頻度を要素とする特徴ベクトルによって表され，その差によって文間の距離は計算される．
4. 文間の距離から類似文を決定する．
5. 類似文の割合や出現位置，距離から文章全体の類似性を評価する．

実験

実験の処理対象はいくつかの真似したレポートが含まれる学生レポートの集合とする．それらのレポートの組み合わせに対し，本手法による類似性評価を行った結果，真似の関係にある文章の組とそうではない文章の組では明確な差が生じた(図 1，図 2)．

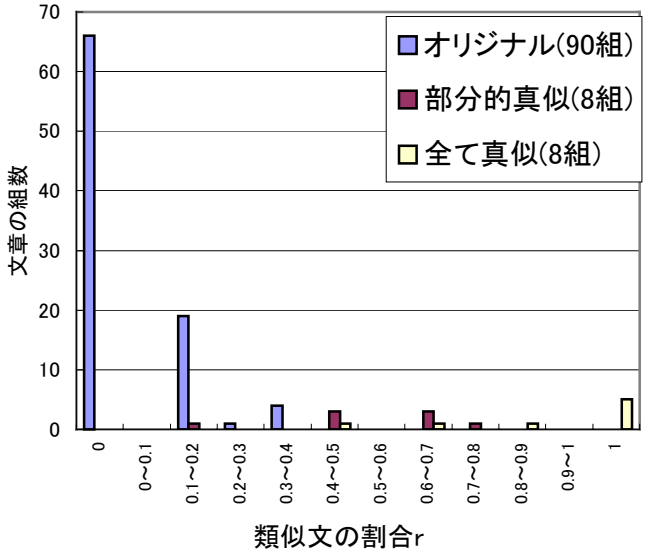

図 1　類似文の割合による比較

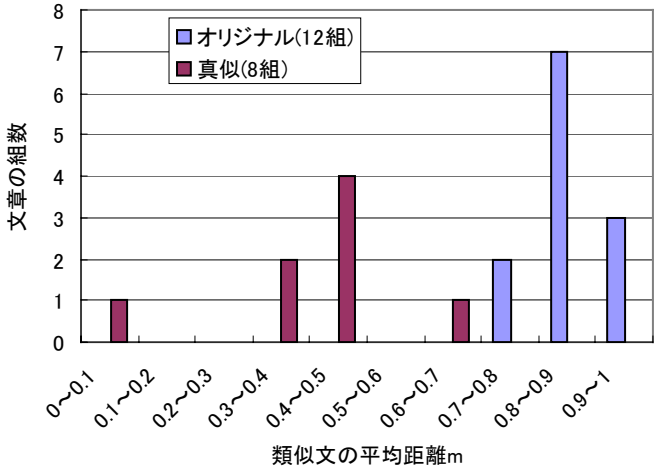

図 2　類似文の平均距離による比較