隣接単語で表現した文脈における高頻度文脈の傾向分析

増山篤志†　梅村恭司†　阿部洋丈†　岡部正幸††

masuyama@ss.ics.tut.ac.jp, umemura@tutics.tut.ac.jp, habe@tut.jp, okabe@imc.tut.ac.jp

† 豊橋技術科学大学 情報工学系　　†† 豊橋技術科学大学 情報メディア基盤センター

概要

文章中に出現する単語と，その前後に出現する隣接単語を文脈と捉えたときの，文脈の出現頻度について分析した．繰り返して出現する文脈の多くは，複合名詞を構成する単語や，定型文として使用される文字列であることがわかった．語の関連性を抽出する技術において，共起する文脈の出現頻度を使用することがあるが，本報告で得られた高頻度の文脈を使用すると，関連性の抽出に悪影響を及ぼす傾向があることを報告する．

1. はじめに

文脈の類似性を扱う研究のひとつに，文脈の類似性から，関連性の高い語を抽出する報告がある[1]．これは，「同じような文脈で出現する語には関連性がある」という仮定の下，文脈の類似性を計算することによって，関連性の高い語を抽出するものである．関連性の高い語としては，類義・同義語や関連語などが挙げられる．これらの語を文章から抽出する技術は，情報検索やシソーラスの構築などの多くの分野に応用できる．

文脈の類似性を求めるための統計値としては，共起する文脈の出現頻度を使用することがある．これは，同じ文脈が何度も共起した場合，文脈の類似性も，より高いものであると考えられるからである．しかし，出現頻度を統計値として使用した場合，高頻度で出現する文脈の影響によって，語の関連性の抽出に悪影響を及ぼす傾向が見られる[2]．これは，tf-idf の考え方と矛盾している．

そこで本報告では，文章中に出現する単語と，その隣接単語を文脈と捉えたときの，文脈の出現頻度と，高頻度出現する文脈の傾向について報告する．

2. 文脈の三つ組み表現

文脈とは，文における個々の語または文の間の論理的な関係や脈絡である．例えば，対象とする語に係る品詞や，近傍に出現する単語や文字列などが挙げられる．実際には，すべての文脈を対象とすることは行われず，文脈の類似性を求める場合，局所的な文脈のみを対象とすることが多く，そのほうが統計値として扱いやすい．

ここでは，あらかじめ文章から単語のみを抽出し，文章中に出現する単語と，単語の前後に出現する隣接単語で三つ組みを作り，三つ組みを文脈として捉える．三つ組みを文脈とした理由は，語の最も近傍に出現する単語との関係性を利用する単純な文脈であり，統計値として比較しやすいためである．

3. 実験

3.1. 実験内容

文章中に出現するすべての単語について三つ組みを作成し，三つ組みの出現頻度を求めた．対象コーパスには NTCIR1 の学術論文記事 20 万件と，毎日新聞 97 年度の記事約 120 万件を使用した．単語の抽出には Mecab を使用し，名詞のみを抽出した．

3.2. 実験結果

作成された三つ組みは NTCIR1 が約 900 万組（約 600 万種類），毎日新聞が約 1200 万組（約 800 万種類）であった．各頻度出現した三つ組みの異なり数の分布を表 1，高頻度出現した三つ組みを表 2 に示す．

4. 考察

表 1 において，どちらのコーパスも，2 回以上出現する三つ組みが，1 回のみ出現する三つ組みの約 1/5 種類という高確率で出現している．今回対象としたコーパスは，論文や新聞記事といった，範囲が限定された文章であるため，同じ分野の文が繰り返して出現しやすい．しかし，1000 回以上も出現している三つ組みもあり，全体的に，繰り返して出現する三つ組みが多すぎる傾向がある．

4.1. NTCIR1 の傾向分析

表 2 の NTCIR1 の三つ組みを見ると，三つ組みのほとんどが，(有限，要素，法)のような，複合名詞を構成する単語であることがわかる．そして，表 2 の三つ組みほどではないが，ある程度頻度の高い三つ組みは，複合名詞を構成する単語が含まれやすい傾向にあった．したがって，複合名詞の出現頻度が，そのまま三つ組みの出現頻度となってしまっている．また，NTCIR1 のような，技術用語が多く出現するコーパスでは，技術用語を組み合わせて複合名詞を作ることが多い．そのため，多くの三つ組みが高頻度で出現したのではないかと考えられる．

複合名詞を構成する単語が，文脈として抽出されてしまうことの是非については，文脈をどう捉えたいのかによる．複合名詞を構成する単語も，論理的な関係によって結合しているため，関係性の抽出をすることができる．しかし，単語の結合性を考えずに，文章中の脈絡や前後関係を文脈として得たい場合，高頻度で出現する文脈は適当な文脈とは言えず，悪影響を及ぼす可能性がある．特に，出現頻度を統計値として使用する場合，表 2 のように，高頻度の文脈の影響が強く出てしまい，文脈の類似性が正しく得られ

表 1　三つ組みの異なり数の分布

出現頻度	三つ組みの異なり数 (NTCIR1)	三つ組みの異なり数 (毎日新聞)
1 回のみ	5131186	6496887
2 回以上	1073442	1476824
10 回以上	49289	55287
100 回以上	1083	2009
1000 回以上	9	28

表 2　高頻度出現した三つ組み

三つ組み (NTCIR1)	頻度	三つ組み (毎日新聞)	頻度
実験, 的, 研究	2109	著作, 権, 交渉	8563
有限, 要素, 法	1962	中, 為, 本文	8561
共, 重合, 体	1511	交渉, 中, 為	8561
基礎, 的, 研究	1501	現在, 著作, 権	8561
有効, 性, 確認	1423	権, 交渉, 中	8561
実験, 的, 検討	1298	為, 本文, 表示	8561
符号, 化, 方式	1273	橋本, 龍太郎, 首相	3841
計算, 機, シミュレーション	1244	日本, 大使, 公邸	3166
導, 電, 性	1005	面, 関連, 記事	2840
明らか, こと, 目的	943	大使, 公邸, 占拠	2745
並列, 計算, 機	934	葬儀, 告別, 式	2486
制御, 系, 設計	812	公邸, 占拠, 事件	2334
光, 導, 波路	810	歳, 葬儀, 告別	2236
幾何, 学, 的	799	ペルー, 日本, 大使	2198
神経, 回路, 網	762	地球, 温暖, 化	1626
導, 波, 管	752	財政, 構造, 改革	1568
梁, 接合, 部	746	小池, 容疑, 者	1380
可能, 性, 検討	733	温暖, 化, 防止	1373
実験, 結果, 報告	716	市, 中央, 区	1252
定量, 的, 評価	716	回, 全国, 高校	1220
温度, 依存, 性	679	日, 米, 防衛	1151
最適, 化, 問題	678	住所, 氏名, 年齢	1149
画像, 符号, 化	674	利益, 供与, 事件	1133
地震, 応答, 解析	651	市, 北, 区	1120
遺伝, 的, アルゴリズム	650	温室, 効果, ガス	1079

なくなってしまう．例えば，「○○を運送する」という文脈から類似性を得ようとしても，同じ文章中に「××運送」という複合名詞が高頻度で出現していた場合，出現頻度の差から，前者の文脈の類似性が弱く捉えられてしまう可能性がある．

複合名詞を抽出しないようにするためには，複合語を分割せずに抽出するツールが必要である．可能性はあるものの，実験では形態素解析器として Mecab を使用し，辞書としては IPA 辞書を使用した．形態素解析器 Chasen でも同様の実験を行ってみたが，ほぼ同様の結果が得られた．状況を改善するために，辞書に複合語の情報を与える方法もあるが，技術用語のような一般的ではない複合名詞を抽出することは難しい．また，一般的ではない複合名詞を抽出する手法としては，統計量によって複合名詞を抽出する手段が考えられる．これには，中川らの提案する専門用語抽出手法などが挙げられる[3]．この方法でも，複合名詞を正確に取ることは難しい問題として残っている．

4.2. 毎日新聞の傾向分析

毎日新聞の三つ組みを見ると，NTCIR1 と同様に複合名詞がいくつか見られる．例えば，（橋本, 龍太郎, 首相）や（公邸, 占拠, 事件）などは，複合名詞として何度も記事に出現したので，高い頻度となっている．

複合名詞以外にも，不自然に頻度が高い三つ組みがある．まず，頻度上位 6 位までの三つ組みは，文章中に「現在著作権交渉中の為，本文は表示できません」という，文章とは無関係なコメントが多数含まれていたために，高い頻度になっていた．これは，ただのノイズになってしまうので，除くべきである．次に，新聞記事の定型文として使用される文字列が多く見られた．例えば，（面, 関連, 記事）の三つ組みは，「3 面に関連記事」という案内表示が多く出現するため，高い頻度になっていた．他にも，（葬儀, 告別, 式）や（市, 中央, 区）なども，定型文として何度も使用されていた．

定型文が，文脈として抽出されてしまうことについては，問題はないと思われる．定型文の文脈からは，人名や土地名など，同じカテゴリに属する単語が抽出できることが期待されるからである（例えば，（県, X, 市）という文脈の類似性から，市の名前が得られる）．ただし，出現頻度を統計値としてそのまま使用する場合，やはり複合名詞や定型文の頻度が高すぎるために，通常の文脈の類似性を正しく得られない可能性がある．

5. まとめ

本報告では，関連性の高い語を抽出する技術で使用されることがある，文脈の出現頻度について分析した．その際，文章中に出現する単語と，前後に出現する隣接単語で三つ組みを作成し，三つ組みを文脈とした．結果，2 回以上出現する三つ組みが非常に多く，特に高頻度の三つ組みは，複合名詞を構成する単語や，特有の定型文として使用される文字列であることがわかった．したがって，文脈の類似性を求めるための統計値として，文脈の出現頻度を使用すると，複合名詞や定型文の頻度が強く影響してしまい，関連性の抽出に悪影響を及ぼす傾向があることを報告する．

6. 参考文献

[1] 當間 雅, 梅村 恭司: 語の出現類似性のための統計的モデルとシソーラス構築への適用, 言語処理学会第 13 年次大会, 2007.

[2] 増山 篤志, 梅村 恭司, 阿部 洋丈, 岡部 正幸: 語の文脈の一致判定における文脈の出現頻度と種類数の比較, 第 193 回自然言語処理研究発表会, 2009.

[3] 中川 裕志, 森 辰則, 湯本 紘彰: 出現頻度と連接頻度に基づく専門用語抽出, 自然言語処理, 19(1), pp.27-45, 2003.