ISS-A-073

潜在意味解析による作成方法の異なる文書の照応

田畑　菜々子　　三浦　孝夫
法政大学理工学部創生科学科

1. 前書き

本研究では、文脈と単語の関係性を捉えるこ
とのできる潜在意味解析を用いて、作成基準の異なる文章間の照応を行う。

2. 照応解析

人間が文章を理解する過程を計算器によって実現するために照応関係を把握することが必要である。そのことを照応解析という。

3. 潜在意味解析

単語単位での情報保持ではなく、類義してそうな単語を 1 つのトピックとし、さらにトピック数に元の「単語-文書行列」を近似する。この近似された共起頻度行列で行相関、列相関を求めるとそれぞれ語相関、列相関が生成され相関係数の値が類似度を示す。

4. 提案手法

本研究では対象を名詞ではなく文章とし、全く異なる表現でも同じ内容を示すという文章間の照応関係を解析できるか否か検証する。そこで潜在意味解析を提案する。

5. 実験

5.1 実験データ

朝日新聞と毎日新聞の 2 週間分（記事数：5952 件、単語数：844 語）用いる。

5.2 実験手順

類義語生成と類似文書作成を行う。新聞記事を、対応付けて相関係数を求め、2 社の新聞記事の照応を行う。最後に、類似性が高いと判断した記事対に関して相関係数が高いか否かを評価する。

5.3 実験結果

表 1 相関係数上位 3 組の類義語生成

1	新党	現	0.99991
2	宗	政	0.999449
3	大臣	現	0.999221

表 2 相関係数上位 3 組の類似文書

	相関係数	毎日新聞	朝日新聞
1	0.967765	M1777	A2610
2	0.8464	M593	A2890
3	0.821844	M1609	A2728

6. 考察

表 1 以外にも「疑い」と「容疑」が生成される。これは類義語であると考えられる。潜在意味解析では共起性を用いて単語をトピック付けができることから、類義語を生成できると考えられる。次に、表 2 では、相関係数が最大の記事対では事実内容が一致し、残りは「逮捕」で一致する。そして、○△においては一致する割合が低くなっている。

表３ 類似性の評価

評価	相関係数	含まれた個数	評価を付けた個数	%
◎	0.85以上	1	1	100
○	0.5以上0.85以下	18	125	14.4
△	0.3以上0.5以下	4	26	15.4

7. 結論

潜在意味解析を用いて、単語の共起性を捉えることができ、それにより類義語を生成することができた。さらに、単語の共起性により単語と文脈を関係づけることができたため、相関係数の値によって作成方法の異なる文章の照応ができた。これによって提案手法の有用性が立証された。

8. 参考文献

[1]石崎 俊(1995)「自然言語処理」
[2]黒崎 禎夫、柴田知秀(2016)「自然言語処理概論」サイエンス社