IDF を利用した n-gram 文字列の分類

下畑さより　　山本秀樹
Sayori SHIMOHATA Hideki YAMAMOTO
沖電気工業(株) 研究開発本部 関西総合研究所
Kansai Laboratory, Research and Development Group, Oki Electric Ind. Co., Ltd.

1 はじめに

近年、コーパスから未知語や専門用語、イディオムなどを自動的に抽出する研究が盛んになっている。なかでも、n-gram 統計を用いる手法は、辞書や形態素解析の使用を前提とせず、コーパスに出現するすべての文字列を文字列長および出現回数を基準に抽出することができる [長尾 93]。一方で、この手法では文字列を網羅的に抽出するため、抽出文字列中に文法的、意味的な単位とならない文字列 (断片的文字列) が多数混在するという問題もある。

これに対して、抽出された n-gram 文字列から断片的文字列を除去する手法が提案されてきた [Smadja93] [池原 95] [下畑 95]。例えば Smadja は、出現回数や出現位置などの制限と簡単な解析処理によって、Subject-Verb、Verb-Object などの共起表現を適合率 80%、再現率 94% で抽出した [Smadja93]。我々も隣接文字の分散の度合 (エントロピー) を基準としてコンピュータマニュアルからの定型表現の抽出を行ない、88.1% の適合率を得ている [下畑 95]。

本稿では、文法的な単位と認められた n-gram 文字列を様々な文書にまたがって出現する文字列 (一般表現) と特定の文書に固有の文字列 (分野依存表現) に分類する方法について述べる。一般表現が様々な分野に偏りなく出現するのに対し、分野依存表現は特定の分野に集中して出現するという特徴に着目して、抽出の基準に情報検索の分野で一般的に用いられる IDF(Inverse Document Frequency) を利用する。また、科学技術論文を使った実験を行ない、本手法の検証を行なうとともに、分類結果の有効な利用方法についても考察する。

[0]連絡先：下畑さより
沖電気工業(株) 研究開発本部関西総合研究所
〒540-6025 大阪市中央区城見 1-2-27 クリスタルタワー
Tel:(06)949-5101,Fax:(06)949-5108,
Email: *sayori@kansai.oki.co.jp*

2 文字列の分類

断片的文字列が既存の手法で除去可能であると仮定すると、n-gram 文字列はその性質から以下の 3 つに分類できる。

1. 一般表現
2. 分野依存表現
3. 1 と 2 の組合せ

一般表現 (GC: General Collocation) は様々な分野に出現する表現で、主に機能語や一般的な名詞で構成されている。GC の意味は広範で、分野に関係なく共通である。GC の種類は限られていてそのうちの多くは中頻度であるが、一部に非常に高頻度なものがある。

これに対して分野依存表現 (DDC: Domain Dependent Collocation) は、特定の分野に頻繁に出現する表現で、そのほとんどは専門用語や固有名詞などの名詞句である。DDC は種類が多く、その意味は分野によって非常に限定されている。

一般型表現と分野依存型表現の組合せ (CGD: Combination of GC and DDC) は、DDC と同様に特定の分野に、ほとんどの場合「名詞句+付属語」の形態で出現する。CDG も種類は非常に多いが、各々の出現頻度は DDC と比べると少ない。

表 1は、コンピュータマニュアルから抽出した n-gram 文字列とその分類を示したものである。

3 TF と IDF

文字列の出現頻度 (TF: Term Frequency) と文書頻度の逆数 (IDF: Inverse Document Frequency) は、ある文字列が特定の文書の索引語として適当かどうかを評価する指標として情報抽出の分野で一般的に用いられている重み付けモデルである [Salton97]。

文字列	分類
on	GC
on the	GC
on the remote host	CGD
remote host	DDC
remote host name	DDC
runtime	DDC
runtime on the remote host	CGD
the	GC

表 1: n-gram 文字列と分類

IDF は、全体の文書数 N における文字列 t_k を含む文書数 n_k の逆数で、以下の式により求められる。

$$idf_k = \log \frac{N}{n_k} \quad (1)$$

また、文書 i における文字列 t_k の評価値 t_{ik} は、文字列 t_k の出現頻度 tf_{ik} と idf_k から、以下の式により求められる。

$$t_{ik} = tf_{ik} \times idf_k \quad (2)$$

つまり t_{ik} は、特定の文書に頻出する文字列ほど大きな値をとり、多くの文書に偏りなく出現する文字列ほど小さな値をとる。

4 TF・IDF を利用した文字列の自動分類

表 1は、本稿で提案する文字列の分類方法の概要である。本手法では、まず複数の文書から抽出した n-gram 文字列から各文書における文字列の TF・IDF を計算し、これを使って GC とそれ以外の文字列を分類する。これは、GC が様々な分野にまんべんなく出現するのに対して DDC と CGD は特定の分野に集中して出現するという特徴を利用したものである。さらに、分類結果を再帰的に利用して、未分類の文字列のうち GC と DDC あるいは CGD とからなる文字列を CGD と分類する。そして、最終的に GC にも CGD にも分類されなかった文字列を DDC と分類する。

GC の分類は、以下の手順で行なう。

1. t_{ik} の計算
2. $t_{ik} \leq T$ であれば t_k を GC に分類
3. t_{ik} が GC で構成されている場合 t_k を GC に分類

まず文書 i 中の各文字列 t_k に対して TF・IDF の値 t_{ik} を計算し、t_{ik} が閾値 T よりも小さい場合、t_k を GC に分類する。次に、未分類の文字列 t_k のうち GC の組み合わせにより構成されているものを t_k そのものの t_{ik} に関わらず GC に分類する。例えば、表 1の文字列において、“on” と “the” が GC に分類されていた場合、“on the” がこの処理によって GC に分類される。この処理は分類結果を再帰的に利用して、条件に当てはまる文字列がなくなるまで繰り返し行なわれる。

次に、未分類の文字列のうち、以下の条件に当てはまる文字列 t_k が CGD に分類される。

4. GC と CGD で構成されている場合、t_k を CGD に分類
5. GC と未分類の文字列 t_i で構成されている場合、t_k を CGD に分類
6. CGD と未分類の文字列 t_i で構成されている場合、t_k を CGD に分類

上記処理も、分類結果を再帰的に利用して、条件に当てはまる文字列がなくなるまで繰り返し行なわれる。そして、最終的に GC にも CGD にも分類されなかった文字列が DDC に分類される。表 1の例では、“on the remote host” は、GC“on the” と未分類の文字列 “remote host” から構成されているため、CGD に分類される。“runtime on the remote host” も同様である。これに対して “remote host” や “remote host name”、“runtime” は、GC や CGD を含む要素で構成されていないので、最終的に DDC に分類される。

5 実験

以上に述べた手順に従い、n-gram 文字列を GC、DDC、CGD の 3 種類に分類する実験を行なった。実験で用いた文書は英日対訳の科学技術記事 47 件、一記事の長さは平均で 204 文である。[下畑 95] の手法に従って断片的文字列を除去した後、分類処理を行なった。

表 2は TF・IDF の閾値 $T = 5$ とした場合の DDC の再現率と適合率を示したものである。再現率は実際の DDC に対して本手法が DDC と分類した文字列の割合、適合率は本手法が DDC と分類した文字列に対して正しい DDC の割合である。再現率と適合率の平均値は各々 0.91、0.89 と、高い値を得ることができた。英語と日本語を比較すると、再現率、適合率ともに英語の方が高くなっているが、これは英語では単語 n-

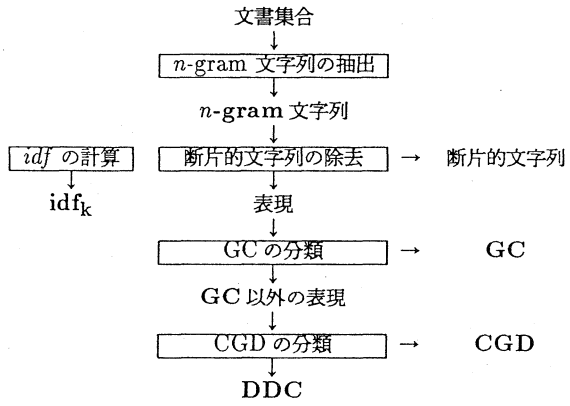

図 1: 処理の概要

gram を使っているのに対し、日本語では文字 n-gram を使っていることが影響していると考えられる。

	～0.85	～0.90	～0.95	～1.00
再現率 (E)	3	2	5	37
再現率 (J)	1	13	15	18
適合率 (E)	8	10	13	16
適合率 (J)	17	11	9	10
				(記事数)

表 2: DDC の再現率と適合率

表 3は、同じ内容の英語と日本語の記事で DDC と分類された上位 10 件の文字列である。分類された文字列はいずれも、記事の分野を反映した専門用語である。また、英語で TF・IDF の高い文字列は、日本語でも TF・IDF が高いという傾向があることがわかった。表 3の場合、10 件中 7 件が両言語に現れている。このことから、本手法により抽出された DDC は、複数言語のコーパスから言語対を抽出する研究 [Kitamura96] [Melamed97] への応用が有効であると考えられる。

表 4は CGD に分類された文字列の例である。これらの文字列は DDC に前置詞、冠詞、助詞などの機能語がついたものがほとんどである。CGD は DDC が文書中でどのような使われ方をするかを示しており、文レベルの分野固有表現の構成要素となることが多い。そのため、非連続な定型表現を抽出する研究 [Shimohata97][Haruno96] に利用することで、抽出対象を絞り込み、必要な表現だけを効率良く抽出することが可能になる。

CGD	TF・IDF	TF
viruses are	30.13	7
infected with	21.52	5
as a moderator	25.82	6
ポーアは	79.44	22
減速材として	38.74	9
スケッチを見た	21.52	5

表 4: CGD と分類された文字列の例

誤分類の多くは、一般的な副詞や形容詞、形容動詞などを DDC と分類したものである。これらの文字列は一般的ではあるがそれほど頻繁には出現しないため、分野依存性が高いと認識された。その他、GC を包含する DDC を CGD とする誤分類も数例あった。表 5は DDC と分類されなかった DDC の例である。例えば、“double star(二重星)” は天文学における専門用語であるが、‘double” が広く使用される語であるため GC“double” と DDC“star” から成る CGD と認識された。この現象は、多用される漢字を含む日本語の場合、特に問題となる。

これらの誤分類は、簡単な解析処理やストップワードリストを導入することで改善が可能である。また、表 5のような文字列は TF・IDF の値が高いことから、CGD を分類する際に TF・IDF の制限を設けることも効果があると考えられる。

DDC	TF・IDF	TF	DDC	TF・IDF	TF
dendrimers	206.60	48	デンドリマー	314.20	73
molecules	96.42	53	分子	115.13	88
monomers	68.87	16	モノマー	75.83	21
methyl acrylate	47.34	11	ポリマー	58.36	20
ethylene diamine	38.74	9	エチレンジアミン	55.95	13
ammonia	36.11	10	アクリル酸メチル	55.95	13
core molecule	34.43	8	水素原子	54.16	15
atoms	34.03	17	合成	52.04	26
core	33.37	15	枝分かれ	47.34	11
polymers	29.64	11	アンモニア	46.94	13

表 3: DDC と分類された文字列の例

文字列	TF・IDF	分類結果
double stars	51.65	CGD
実無限	43.04	CGD
化学的	13.47	GC

表 5: 分類に失敗した文字列の例

6 まとめ

TF・IDF を用いて文字列を一般表現、分野依存表現、一般表現と分野依存表現の組み合わせの 3 つに分類する方法について述べた。また科学技術記事を使った実験を行ない、n-gram 文字列中の分野依存表現を再現率 91%、適合率 89% で分類できることを示した。

n-gram 文字列は、何らかの形で他の自然言語処理アプリケーションに適用されることが前提になっている。本手法は、n-gram 文字列を分類することでアプリケーションが必要な n-gram 文字列だけを利用できるようにすることを目指している。これは、対象を絞り込むことにもなるので、処理効率の面からも有効である。今後は、考察で述べたように分類結果を他の言語処理技術に適用し、本手法の実用的な効果を検証する。

参考文献

[Haruno96] Haruno,M., Ikehara,S., and Yamazaki,T.: Learning Bilingual Collocations by Word-Level Sorting, COLING96, pp.525-530(1996).

[池原 95] 池原, 白井, 河岡: 大規模日本語コーパスからの連鎖型および離散型の共起表現の自動抽出法, 情報処理学会論文誌 Vol.34, No.9, pp.1937-1943(1995).

[Kitamura96] Kitamura, M. and Matsumoto, Y.: Automatic extraction of word sequence correspondences in parallel corpora, Proceedings of 4th WVLC, pp.79–87(1996).

[Melamed97] Melamed,I.D.: A Word-to-Word Model of Translational Equivalence, Proceedings of the 35th Annual Meeting of ACL, pp.490–497(1997).

[長尾 93] 長尾, 森: 大規模日本語テキストの n グラム統計の作り方と語句の自動抽出, 情報処理学会自然言語処理研究会報告 96-1, pp.1–8(1993).

[下畑 95] 下畑, 杉尾, 永田: 隣接文字の分散値を用いた定型表現の自動抽出, 情報処理学会自然言語処理研究会 110-11 pp.71-78(1995).

[Shimohata97] Shimohata,S., Sugio,T., and Nagata,J.: Retrieving Collocations by Co-occurrences and Word Order Constraints, Proceedings of the 35th Annual Meeting of ACL, pp.476–481(1997).

[Salton97] Salton, G. and Buckley, C.: Term Weighting Approaches in Automatic Text Retrieval, Readings in Information Retrieval, edited by Sparck Jones, K. and Willett, P. Morgan Kaufmann, pp.323–328(1997).

[Smadja93] Smadja,F.: Retrieving Collocations from Text: Xtract, Computational Linguistics, Vol.19, No.1, pp.143-177(1993).

[Smadja96] Smadja,F.A., MaKeown,K. and Hatzivassiloglou, V.: Translating Collocations for Bilingual Lexicon: A Statistical Approach, Computational Linguistics, 22(1), pp.1–38(1996).