Appendix. 7　特徴語の抽出

7-1　研究領域における特徴語とは

サイエンスマップにおいて、各研究領域の内容を把握することは重要である。しかし、各研究領域に含まれるコアペーパやサイティングペーパに全て目を通すことは困難である。そこで、サイエンスマップでは各領域にその領域の特徴を表す語（以下、特徴語）を複数個付与している。サイエンスマップ 2014 では文書中の単語に関する重み付けを行う手法である Term Frequency・Inverse Document Frequency（TF・IDF)を用いて外部機関が特徴語群を抽出し、NISTEP 内で目視確認を行った。これに対して、サイエンスマップ 2016 では科学技術振興機構 情報企画部と共同で、BM25 手法[1]をベースに次節に示す 4 つの特性（代表性、共通性、特定性、抽象性）に合わせて抽出された 4 つの特徴語セットを組み合わせて構成する方法を開発した。

7-2　特性に合わせた特徴語抽出手法の提案

本節では、まず特徴語の特性に関する定義を行った後、それぞれの特性に特化した特徴語抽出手法を説明する。

7-2-1　特性の定義

サイエンスマップに付けられるべき特徴語の特性として、徳永ら[2]の研究をベースに以下の 4 項目を定義した。

○　代表性

領域内に含まれる研究テーマの内、もっともその領域を代表すると思われるテーマ、およびそのテーマを表す特徴語。端的な表現としては、その領域内でもっとも目立つテーマを意味する。例えば、10 論文の内、7 論文が膠原病の一種、全身性エリテマトーデスに関する論文であれば、「全身性エリテマトーデス」が代表語である。先の TF・IDF の値が高くなる語は、主に代表性の高い語である。

○　共通性

領域内に含まれる研究テーマ群の中で、もっとも多くのテーマに共通するテーマ、およびそのテーマを表す特徴語。代表性と必ずしも相反するわけではないが、多くの場合、代表性よりも抽象的な概念であったり、基礎的な技術を指す場合が多い。また、代表性と異なり、共通性を重視すると他領域との違いが明確でなくなるという問題も指摘される。例えば、上記の例において残り 3 つが全身性エリテマトーデス以外の膠原病に関する論文であった場合、「膠原病」が共通語である。

○　特定性と抽象性

代表性－共通性と直行する軸として、領域内の位置づけではなく、テーマや特徴語自身が表す概念の粒度を特定性と抽象性として表す。例えば、固有の細胞名や技術的手法などは特定性が高く、複数のそれらをまとめた臓器の名称や技術的目的・課題などは抽象性が高いものとして定義する。階層的に整理されたシソーラスにおいては、上位語ほど抽象性が高く、下位語ほど特定性が高い。オントロジーにおいては、is-a や part-of 関係における上位語（広義語）と下位語（狭義語）である。また、代表性と特定性、共通性と抽象性を表す語は必然的に相関が強いものと想定される。

事前調査として、サイエンスマップ 2014 に付けられていた特徴語を精査したところ、TF・IDF 手法の性質から代表性と特定性を表す語が多いことが分かった。そこで、サイエンスマップ 2016 では、各特性をニーズに合わ

せてバランスさせた特徴語セットを抽出することを目指す。但し、サイエンスマップ 2016 では抽象性は使用していない。

また、事前調査の結果、一部に科学技術用語として相応しくない特徴語が含まれていたことも分かった。多くは平易な形容表現や日付、地名などである。そこで、これらを除外するため、特徴語を JST 科学技術用語シソーラス・大規模辞書[3]に含まれる約 130 万語（同義語含む）、および元論文に付与されたキーワードに限定することとした。

7-2-2　特徴語抽出手法（ベースライン）

昨今、自然言語処理におけるトピック抽出手法としては Latent Dirichlet Allocation (LDA)[4]がしばしば用いられる。これはトピック群の確率がディリクレ分布に従うことを仮定して、もっともよく説明するトピック群を確率的に求めるものである。しかし、今回は後述する 3 つの特性に沿う特徴語群を構成し、それらを一定の割合で組み合わせてサイエンスマップの利用者に適した特徴語セットとすることを目的としているため、確率計算過程のチューニングが困難なLDAは使用しなかった。また、一定の文書群から3層のニューラルネットワークモデルを用いて、単語ベクトルや文書ベクトルを構築する手法 word2vec[5], doc2vec[6]も注目を集めている。例えば、doc2vec で構築されたベクトルを対象に、事前にラベリングされた文書－特徴語群を教師データとして与えて教師付き学習させることで、いわば分類学習として文書群への自動ラベリングを行うことができる。しかし、今回は領域毎に異なる特徴語群を付けるものであり、分類学習として十分な教師データを用意することが難しいため、ベクトル表現は使用が困難であった。そこで今回は、前回までの TF・IDF 手法を一般化した BM25 手法を用いて、サイエンスマップ各領域から特性毎の特徴語を抽出した。

サイエンスマップ 2014 で用いた TF・IDF 手法は、文書中の単語に関する重みの一種であり、情報検索や文章要約などの分野で経験的に有意であることが広く知られている。しかし、これはヒューリスティックな手法であり、必ずしもサイエンスマップ各領域のラベリングに適しているとは言えない。そこで、以下ではまず TF・IDF 手法を平均相互情報量の特殊な形式であることを示し[7]、特殊化された際の条件がサイエンスマップの領域ラベリングには不向きであることを示し、より一般的な形式をベースライン手法として示す。

確率論および情報理論における平均相互情報量 I とは、2 つの確率変数の相互依存の尺度を表す量であり、単語集合と文書集合を表す 2 つの確率変数 W と D の平均相互情報量は以下で定義される。

$$\begin{aligned} I(W;D) &= H(W) - H(W|D) \\ &= \sum_{w_i \in W} \sum_{d_j \in D} p(w_i, d_j) \log \frac{p(w_i, d_j)}{p(w_i)p(d_j)} \end{aligned} \tag{1}$$

ここで $H(W)$ は選択情報量（自己エントロピー）、$H(W|D)$ は条件付きエントロピーを表し、$p(w_i,d_j)$ は W と D の同時生起確率である。直観的には、平均相互情報量 I は W と D が共有する情報量を表す。一方、自己エントロピー $H(W)$ は確率変数の不確かさ（単語の出現の偏り）の尺度であり、$H(W|D)$ は D を指定した後にも残る W の不確かさであるため、I は D が指定されたことで削減される W の不確かさであるとも言える。ここで簡略化のため、それぞれの確率を以下のように定義すると、

$$
\begin{aligned}
p(w_i) &= \frac{f_{w_i}}{F} \\
p(d_j) &= \frac{f_{d_j}}{F} \\
p(w_i, d_j) &= \frac{f_{ij}}{F}
\end{aligned}
\tag{2}
$$

式(1)は式(3)のように展開される。尚、f_{ij} は文書 d_j に含まれる単語 w_i の出現頻度、f_{wj} は w_i の全文書分の総和、f_{dj} は文書 d_j に含まれる単語数、F は全文書に含まれる単語数を表す。

$$
\begin{aligned}
I(w_i; d_j) &= \frac{f_{ij}}{F} \log \frac{\frac{f_{ij}}{F}}{\frac{f_{d_j}}{F} \cdot \frac{f_{d_j}}{F}} \\
&= \frac{f_{ij}}{F} \log \frac{\frac{f_{ij}}{f_{w_i}}}{\frac{f_{d_j}}{F}}
\end{aligned}
\tag{3}
$$

式(3)において以下の 2 条件が成立する場合、

$$
\begin{aligned}
\frac{f_{d_j}}{F} &\approx \frac{1}{|D|} \\
\frac{f_{ij}}{f_{w_i}} &\approx \frac{1}{|D_i|}
\end{aligned}
\tag{4}
$$

式(3)は式(5)に展開される。$|D|$ は全文書数、$|D_i|$ は単語 w_i を含む文書数であり、この式(5)が TF･IDF に相当している。

$$
I(w_i; d_j) \approx \frac{f_{ij}}{F} \log \frac{|D|}{|D_i|}
\tag{5}
$$

つまり、式(4)に表された 2 つの条件、

- 各文書の語数はほぼ同じである。
- 特定の単語を含む文書間におけるその単語の出現頻度はほぼ同じである。

が成立する場合、平均相互情報量 $I(w_i,d_j)$ は TF･IDF を表す。言い換えれば、TF･IDF は先の 2 条件下での平均相互情報量の特殊な形式であると言える。

しかし、サイエンスマップの各領域に含まれる文書数（論文数）には大きな偏りがあり（数 10～1000 程度）、特定の技術用語の出現頻度も領域間を跨いで均一ではない。したがって、今回の問題に TF･IDF をそのまま用いることはあまり適切ではないと考えられる。

一方、これまで TF･IDF にはさまざまなヒューリスティックな改良が施されており、その中には精度が高いことでよく知られた BM25[1]が挙げられる（現在、世界中でもっともよく使われていると言われる、フリーの検索エンジン Apache Solr でも 2016 年にリリースされた Ver.6 より、BM25 がデフォルト実装となっている）。

$$
BM25(w_i, d_j) = \frac{f_{ij} \cdot (k_1 + 1)}{f_{ij} + k_1 \cdot \left(1 - b + b \cdot \frac{F_j}{F_{ave}}\right)} \log \frac{|D| - |D_i| + 0.5}{|D_i| + 0.5}
\tag{6}
$$

ここで、F_{ave} は各文書に含まれる単語数の平均、F_j は文書 d_j に含まれる単語数を表し、パラメータは $k_1 = 2.0$、$b = 0.75$ がしばしば用いられている。

そこで、本稿では式(6)を式(4)で表された一般的な形式に戻した式(7)をベースライン手法とする。そして、確率 $p(w_i,d_j)$（ベースライン手法では式(2)に沿って f_{ij}/F と定義．以下では $1/F$ は正規化定数として省略する）によるモデリングを変更することによって、先に定義した特性に沿った指標とする。

$$BM25K(w_i,d_j) = \frac{f_{ij}\cdot(k_1+1)}{f_{ij}+k_1\cdot\left(1-b+b\cdot\frac{F_j}{F_{ave}}\right)}\log\frac{\frac{F}{f_{d_j}}-\frac{f_{w_i}}{f_{ij}}+0.5}{\frac{f_{w_i}}{f_{ij}}+0.5} \tag{7}$$

7-2-3　特徴語抽出手法（代表性）

前述したように代表性は、文書内でよく見かけることでいわば強く印象づけられる研究テーマ、用語である。そこで単語 w_i と文書 d_j の同時生起確率 $p(w_i,d_j)$ の分布として、文書内に含まれる単語の出現頻度 f_{ij} を定数 k 倍した分布を仮定する（実際には $k = 2$ と設定した）。

$$p(w_i,d_j) = kf_{ij} \tag{8}$$

但し、単語 w_i の出現頻度 f_{ij} は文単位でカウントし、同一論文のタイトル、抄録、キーワードにおいて制限なくカウントするものとする。更に、ヒューリスティックスとして論文に付与されたキーワードは論文のトピックをよく表すものであるため、キーワードに現れた場合は出現回数を $+n$ 回とする（実際には $n = 3$ と設定した）。

7-2-4　特徴語抽出手法（共通性）

一方で、共通性は領域内の多くの論文に共通して現れる研究テーマ、用語である。そのため、1 論文内の単語の出現回数は最大 1 回までとし、キーワードによる出現回数のバイアスも行わないものとする。

$$p(w_i,d_j) = \frac{\sum_{l\in d_j}E_{il}}{f_{d_j}}$$
$$E_{il} = \begin{cases}1 & (f_{il}>0)\\ 0 & (otherwise)\end{cases} \tag{9}$$

ここで、E_{il} は単語 w_i が論文 l のタイトル、抄録、キーワード、雑誌名のいずれかに含まれていれば 1 を、そうでなければ 0 を返す関数である。代表性との直感的な違いは、出現頻度を文単位でカウントするか論文単位でカウントするかであると言える。それぞれ文章 d_j における高頻度語重視と共通語重視に相当している。

7-2-5　特徴語抽出手法（特定性、抽象性）

最後に、単語の特定性、抽象性を測るために、平均情報量（エントロピー）$H(C)$ を導入する。エントロピーは、特定の概念（代表語とその同義語）、およびその概念の JST シソーラス内での下位概念（代表語とその同義語）の文書内での出現の偏りを表し、言い換えればその概念の意味的な広がりを表している。ここでは幅広い文脈でさまざまな言い方で使われる概念は意味的な広がりが大きく、抽象性が高いものと仮定している。反対に、特定の文脈で同じ表現でのみ使われる概念は特定性が高いものと仮定する。

図 1 および式(10)に表すように、特定の概念 C のエントロピーは上位語 T_0 とその下位語 $T_1 \ldots T_n$ の出現

頻度を各事象確率と捉えることで計算する。各概念 T_i の同義語 $S_{i0}...S_{im}$ の出現頻度は、対応する語に合算される。ここで $p(S_{ij}|C)$ は、概念 C と語 T_i が与えられた際の同義語 S_{ij} の出現確率を表す。以下の実験では、サイエンスマップ 2016 の全論文のタイトル、抄録、キーワードにおける JST シソーラス内の各概念のエントロピー $H(C)$ を事前に計算している。データセット内で、事象（上位概念／下位概念の同義語）が等しく出現するほどエントロピー $H(C)$ は増大し、特定の同義語のみが現れる場合は（起こりやすい事象の情報量は低いため）エントロピーは低くなる。このようにエントロピーの程度は概念の意味的な広がりを表している。

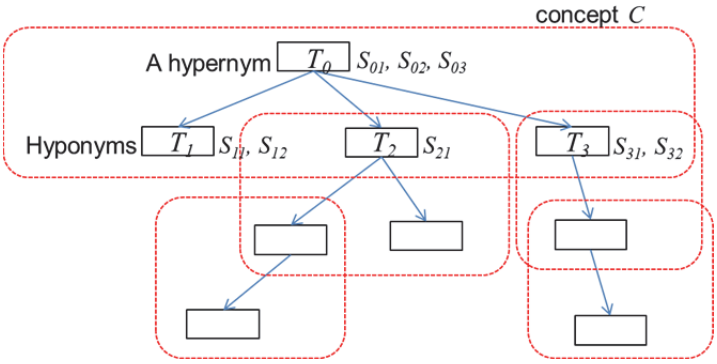

図 1. JST シソーラス内の概念構造

但し、下位概念を持たない階層構造上で末端の概念や、1 種類の同義語のみが現れる概念に関しては、自動的に特定性が高いものと判断し、$H(C) = 0$ とする。論文から取得したキーワード群も同様である。また、定数 q は、頻度とエントロピーのオーダーを合わせるための係数である（実際には $q = 10$ と設定した）。尚、C_{wi} は単語 w_i が属する概念を表す。

$$p(w_i, d_j) = \begin{cases} (1 + qH(C_{w_i})) \cdot f_{ij} & \text{（抽象性）} \\ f_{ij}/(1 + qH(C_{w_i})) & \text{（特定性）} \end{cases}$$

$$H(C) = -\sum_{i=0}^{n}\left(\sum_{j=0}^{m} p(S_{ij}|C) \cdot \log_2 \sum_{j=0}^{m} p(S_{ij}|C)\right) \quad (10)$$

7-3　特徴語抽出処理

特徴語の抽出においては、研究領域を構成する全ての論文（コアペーパおよびサイティングペーパ）を用いた。具体的には、サイエンスマップ 2016 における各領域に含まれる論文のタイトル、抄録、キーワードの 3 項目を対象に、ベースライン手法（式(7)）を特性に合わせて調整した式(8)、式(9)、式(10)（上・下）によって各 10 語、4 種類の特徴語セットを出力した（重複を含む）。また、特徴語を JST シソーラス・大規模辞書用語またはキーワードに限定するため、文字列を Stanford coreNLP（https://stanfordnlp.github.io/CoreNLP/）で語幹化、小文字化した上で、シソーラス用語またはキーワードに完全一致した語のみをカウントしている（編集距離 Levenshtein distance で、スコア 0.9 以上で一致した語をカウントした場合と大きく変わらないことを確認している）。また、JST シソーラスは特定の概念を表す代表語とその同義語で構成されているため、1領域において代表名とその同義語のいずれか複数が出現した場合は代表名［同義語 1，同義語 2,...］としてまとめて表記し、いずれか1つのみが出現した場合は単名で表記した。なお、略語とフルスペルはカウントを合算している。さらに、特徴語セット内において他の特徴語に包含され（部分文字列となり）、かつ、カウントの差が 10%以下である語は省略した。このように抽出された特性毎の特徴語を目視確認した結果、例えば特徴語を 10 語出力した場

合、7 個前後は指定の特性に沿った特徴語となっていることを確認した。その他に、ストップワードリスト（除外語）リストを用意し、不適切な語は適宜登録、除外した。

最後に、ここまでの特徴語は英語で書かれた論文から抽出しているため英語表記となっている。そこで、特徴語から研究領域の内容を把握しやすくするため、特徴語の和訳を行った。翻訳にあたっては、JST シソーラス・大規模辞書由来の特徴語はシソーラスに定義された対訳語を付与した。また、過去のサイエンスマップ作成に日本語に訳されたことのある語は、その際に作成した対訳辞書を使用した。それ以外の新規の特徴語に関しては、該特徴語を含む文を全論文からランダムに 10 文抽出し、文全体を Google 翻訳エンジンを用いて日本語に翻訳し、特徴語に対応する日本語文字列を抽出し、最も多い訳語を選択した。最後に、全訳語の目視確認を行った。報告書に掲載されているのは、和訳された特徴語である。但し、これらは報告書執筆者による仮訳であり、より適切な和訳が存在する可能性がある点について留意願いたい。

また、最先端の研究でかつ国際共同研究が主に行われている研究領域については、和訳を見出すことが困難な特徴語が多数みられた。これらの特徴語については、英語の特徴語をそのまま掲載している。ある英語の科学技術用語に対応する日本語が存在するかについては、我が国における研究者コミュニティーの有無、研究者コミュニティーの大きさ、研究の進展の速度、研究の国際化の度合、科学研究と社会とのつながりの度合などが関係していると思われる。しかしながら、この仮説の検証は、本調査研究の範囲を超えることから、今後の研究の進展を待つこととしたい。

参考文献

[1] E. Garcia: “A Tutorial on OKAPI BM25 Model,” http://www.minerazzi.com/tutorials/okapi-bm25-model.pdf, 2016.

[2] 徳永健伸: “情報検索と言語処理,” 東京大学出版会, 1999.

[3] 川村隆浩, 渡邊勝太郎, 松邑勝治, 櫛田達矢, 古崎晃司: “JST 科学技術用語シソーラスの Linked Data 化：科学技術情報をリンクする知識インフラの構築に向けて,” 情報管理, 59(12), pp. 839-848, 2016.

[4] Blei, D.M., Ng, A.Y., and Jordan, M.I.: “Latent Dirichlet Allocation,” Journal of Machine Learning Research, 3, pp.993-1022, 2003.

[5] Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S. and Dean, J.: “Distributed Representations of Words and Phrases and their Compositionality,” Proceedings of the 26th International Conference on Neural Information Processing Systems, pp.3111-3119, 2013.

[6] Le, Q. and Mikolov, T.: “Distributed Representations of Sentences and Documents,” Proceedings of the 31st International Conference on Machine Learning, 32, 2014.

[7] A. Aizawa: “The Feature Quantity An Information Theoretic Perspective of Tfidf-like Measures,” Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval, pp. 104-111, 2000.