共起情報に基づく医学文献からの略語抽出

岡崎　直観 †§　　Sophia Ananiadou ‡§　　辻井　潤一 †§

† 東京大学大学院情報理工学系研究科
‡ School of Informatics, University of Manchester, UK
§ National Centre for Text Mining, UK

1　はじめに

生物・医学分野では，遺伝子，たんぱく質，化学化合物，薬品，有機体などで使用される用語の数が，飛躍的に増加している．既存の用語資源や科学データベース（例えば Swiss-Prot[1]，SGD[2]，FlyBase[3]，UniProt[4] ）は，この増加のスピードに対応できず [11]，生物・医学分野の文献を扱う上で，用語の管理は重要な役割を担っている [2, 4, 7] ．用語の管理には様々なタスクがあるが，綴り，語形，類義語，概念階層等によるバリエーションの認識は，主要なタスクの一つである．

略語は文献の著者により活発に生成される用語のバリエーションである．一般的に，略語（例えば *RARA*）は，語（例えば *retinoic acid receptor alpha*）を短縮したものであり，しばしば元の語の代用として単独で用いられる．本稿において，略語は完全形（long form）中の文字を抜き出し，短縮形（short form）にしたものと定義する．Wren ら [14] は，*c-jun N-terminal kinase (JNK)* に関する文献を PubMed[5] で検索する際，完全形である *c-jun N-terminal kinase* に適合する文書の数は 3,773 件であったのに対し，短縮形である *JNK* に適合する文書は 5,477 件まで拡大されると報告している[6]．Chang ら [5] は，生物・医学分野の文献において 2004 年に約 64,242 件の略語が新たに生成されたと報告している．

このように，生物・医学文献から略語の短縮形と完全形との対応関係を発見することは，用語管理の重要な課題となっている．生物・医学文献からの自動略語抽出に向けて，抽出ルールに基づく手法 [3, 8, 9, 12]，短縮形と完全形との文字走査に基づく手法 [1, 13]，学習（線形回帰）に基づくアプローチ [5] が提案されている．これらの手法は，略語の短縮形と完全形の文字並びの一致度合いに基づくため，*thyroid transcription factor-1 (TTF-1)* のように文字のアライメントに曖昧性がある場合，*acquired immunological disease (AIDS)* のように略語を正しく定義していない場合，*beta2 adrenergic receptor (ADRB2)* のように文字の並び順が短縮形と完全形で異なる場合，略語抽出に失敗したり，複雑なルール記述が必要となる．

生物・医学分野では，PubMed に代表される文献データベースが整備されており，大規模なテキストコーパスが入手可能な状況にある．そこで，本研究では大規模なテキストコーパス中の略語の短縮形と完全形の共起情報に着目し，略語辞書の自動構築に向けた手法を提案する．提案手法は，略語の短縮形と完全形の文字の並びに依存せず，短縮形・完全形ペアのスコア（確からしさ）を推定可能という特徴がある．本稿ではさらに，完全形のスコアを利用して，略語辞書に登録する上で冗長な完全形（例えば *central nervous system injury (CNS)*）を縮退させる方法，文字走査に基づく従来手法 [13] との組み合わせについても述べる．なお，本稿では，提案手法を英語の文献に適用することと想定する．

2　共起情報に基づく略語抽出

図 1 に略語抽出システムの概要を示す．本研究における略語抽出問題は，大規模な入力テキストが与えられたとき，その中に含まれる略語の完全形と短縮形のペアを，すべて列挙することと定義する．この問題を，以下の 4 つのステップに分割する．

1. **短縮形認識**: 入力文書中に含まれる略語の短縮形を全て認識・列挙する
2. **文蓄積**: 略語の短縮形を含む文の全てをデータベースに蓄積する

[1] http://www.ebi.ac.uk/swissprot/
[2] http://www.yeastgenome.org/
[3] http://www.flybase.org/
[4] http://www.ebi.ac.uk/GOA/
[5] http://www.ncbi.nlm.nih.gov/entrez/
[6] 2004 年 5 月 14 日時点

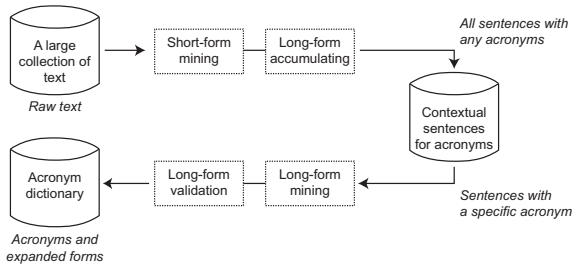

図 1: 略語抽出システムの概要

表 1: 略語を含む文の例

略語	その略語を含む文
...	....
HML	Hard metal lung diseases *(HML)* are rare, and complex to diagnose.
HMM	Heavy meromyosin *(HMM)* from conditioned hearts had a higher Ca++-ATPase activity than from controls.
HMM	Heavy meromyosin *(HMM)* and myosin subfragment 1 (S1) were prepared from myosin by using low concentrations of alpha-chymotrypsin.
HMM	Hidden Markov model *(HMM)* techniques are used to model families of biological sequences.
HMM	Hexamethylmelamine *(HMM)* is a cytotoxic agent demonstrated to have broad antitumor activity.
HMN	Hereditary metabolic neuropathies *(HMN)* are marked by inherited enzyme or other metabolic defects.
...	... .

3. **完全形認識**: それぞれの略語の短縮形に対し，その略語を含む全ての文に用語抽出手法を適用し，完全形の候補とスコアを作成する

4. **完全形抽出**: 完全形の候補に対して，短縮形との対応関係を検証するとともに，冗長・不要な完全形候補を削減する．

最初のステップ（短縮形認識）では，入力文章の中で略語と思われる箇所を認識する．従来研究のほとんどは，括弧表現を手がかりとしたパターンを使って短縮形を抽出している．

long form '(' short form ')'

本研究では，Schwartz ら [13] の手法と同様に，括弧内のフレーズが,「2 語以下」,「2〜10 文字で構成」,「1 文字以上のアルファベットを含む」,「先頭の文字がアルファベットまたは数字」という条件を満たす場合のみ，そのフレーズを略語の短縮形として認識する．

第 2 のステップ（文蓄積）では，略語の短縮形を含む全ての文がデータベースに追加される（表 1 の例を参照）．後で詳しく述べるように，提案手法は特定の略語を含む文集合に含まれる共起情報を分析することによって，完全形の候補を生成する．このデータベースは，略語の短縮形が与えられると，その短縮形を含む文のみを全て返すように設計される．

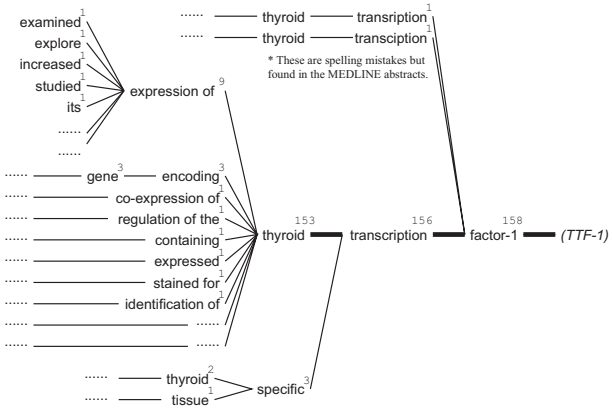

図 2: 略語 *TTF-1* の完全形候補とその共起頻度

第 3 のステップ（完全形認識）は，提案手法の核心部分である．今までの手順で認識された個々の短縮形を含む文を分析し，短縮形とよく共起する（複合）語に着目することで，略語に対応する完全形の候補リストを作成する．図 2 は，略語 *TTF-1* を含む文を全て分析し，その略語の直前に存在していた語を，木構造で示したものである．任意のノードからノード *factor-1* に到る経路上に存在する語を並べると，略語 *TTF-1* に対する完全形の候補が完成する．なお，完全形の候補は，非機能語[7]を先頭に，短縮形の直前の語（図 2 の場合は *factor-1*）までを省略なく並べたものと定義する．完全形の候補と略語 *TTF-1* との共起回数は，出発ノードの右上に示されている．図 2 は，完全形候補 *factor-1*，*transcription factor-1*，*thyroid transcription factor-1* が，略語 *TTF-1* とそれぞれ 158，156，153 回共起していることを示している．

完全形候補 *factor-1* や *transcription factor-1* は，短縮形 *TTF-1* と頻繁に共起しているが，これらの候補は周辺語 *thyroid* とも頻繁に共起している．これに対し，候補 *thyroid transcription factor-1* は，様々な文脈（例えば，*expression of thyroid transcription factor-1*, *expressed thyroid transcription factor-1*, *gene encoding thyroid transcription factor-1* など）で用いられているため，短縮形 *TTF-1* と候補 *thyroid transcription factor-1* の間に，何らかの関係があると推測される．この場合は，候補 *thyroid transcription factor-1* が短縮形 *TTF-1* の文字のすべてを含むため，略語の短縮形と完全形の関係にあると推定される．

[7]略語抽出システムには，*a, the, of, be* などの機能語が 133 件登録されている

以上の議論から，完全形候補として注目すべき語は，略語の短縮形と頻繁に共起するが，それ以外の周辺の語とは共起しない語と仮定する．略語の短縮形を含む文から完全形候補を抽出する問題は，与えられたテキスト（この場合はある略語の短縮形を含むすべての文）から，用語を自動抽出する問題に帰着される．C-Value 法 [6] は，与えられたテキストの中で頻発に出現し，かつ他の語の一部として使われることが少ない複合語を抽出する．C-Value 法の詳細については割愛するが，品詞情報に基づいて抽出した語の候補に，出現頻度とネスティング関係（語の包含関係）を用いて，単語らしさのスコア付けを行うものである．オリジナルの C-Value 法は，品詞情報を必要とする，２単語以上から構成される複合語を抽出する，単語数の多い語を優先するなど，そのままでは完全形候補のスコア付け法として利用できないので，以下のように改良した．

ある略語の短縮形を含む文をすべて集め，図 2 の例で示した完全形の候補をすべて列挙するため，以下のフィルタを適用する

```
[:WORD:].*[:WORD:]$
```

ここで，`[:WORD:]` は任意の非機能語，`.*`は 0 語以上の任意の語，`$`は略語の短縮形にマッチする．なお，語は Porter のステミング・アルゴリズム [10] を適用し，標準的な形へ統一しておく．

このように列挙された語 w に対し，以下の式で定義される単語スコア $TH(w)$ を計算する．

$$\mathrm{TH}(w) = \mathrm{freq}(w) - \sum_{t \in T_w} \mathrm{freq}(t) \times \frac{\mathrm{freq}(t)}{\mathrm{freq}(T_w)}. \quad (1)$$

ここで，w はスコア付けを行う完全形候補，$\mathrm{freq}(w)$ は候補 w の出現頻度，T_w は候補 w に一語を追加して生成される完全形候補の集合，$\mathrm{freq}(T_w)$ は T_w に含まれる完全形候補の出現頻度の総和である．式 1 の第 1 項は，略語の短縮形と完全形候補 w の共起回数そのものである．第 2 項は，完全形候補 w の共起回数の中で，w に任意の 1 語を追加して生成される完成形候補が占めている共起回数の期待値を減じるものである [8]．

表 2 は，略語 *TTF-1* と頻繁に共起する候補語 w を，スコア $\mathrm{TH}(w)$ の高い順に並べたものである．語 *factor-1* や語 *transcription factor-1* は共起頻度こそ高いものの，

表 2: 略語 *TTF-1* と頻繁に共起する語のリスト

略語の完全形候補 w	$\mathrm{freq}(w)$	$TH(w)$
thyroid transcription factor-1	153	150.7
expression of thyroid transcription factor-1	9	7.77
transcription factor thyroid transcription factor-1	7	6.0
transcription factor-1	156	5.9
factor-1	158	4.0
factor thyroid transcription factor-1	8	1.8
gene encoding thyroid transcription factor-1	3	1.3
co-expression of thyroid transcription factor-1	1	1.0
regulation of the thyroid transcription factor-1	1	1.0
....	...	...

その共起頻度の殆どはそれぞれ，語 *transcription factor-1* と語 *thyroid transcription factor-1* に起因するものであるため，式 1 の第 2 項で共起頻度の大半を失う．これに対し，語 *thyroid transcription factor-1* は，*expression of thyroid transcription factor-1* などの語に含まれるものの，T_w には様々な種類の語が存在するため，T_w が占める頻度の期待値は小さく，共起頻度の殆どがスコア $TH(w)$ として残る．

図 1 の第 4 ステップ（完全形抽出）は，完全形候補とそのスコアのリストを使って，最終的に抽出する完全形を決定する．今回は，略語の短縮形と完全形のペアのみを抽出したい[9]ため，短縮形のすべての文字を完全形が含むか確認する．表 2 によると，短縮形 *TTF-1* の完全形として最も適当な語は *thyroid transcription factor-1* である．そこで，完全形 *thyroid transcription factor-1* から略語 *TTF-1* が生成されたと仮定すると，その完全形の部分的に含む語（*transcription factor-1* と *factor-1*）や，完全形に別の語を追加して生成される語（*expression of thyroid transcription factor-1* など）は，短縮形を生成させるための要素が欠けていたり，余分な要素が付け足されていると考えられる．そこで，「すでに抽出された完全形に対し，語を追加もしくは削除して得られる完全形候補は抽出しない」というルールを導入し，スコア $\mathrm{TH}(w)$ の高いものから順に閾値 θ 以上の完全形候補を抽出し，短縮形・完全形のペアを確定させる．表 2 に示されている完全形候補に対して，この抽出アルゴリズムを適用すると，最終的に抽出される完全形は *thyroid transcription factor-1* のみとなり，その他の候補はすべて削除される．

3 評価

生物・医学分野からの略語抽出の評価セットとして，Medstract [12] の正解データがよく用いられるが，正解

[8] オリジナルの C-Value スコアは，第 1 項の $\mathrm{freq}(w)$ に候補語 w の語数の対数を掛けて，語数の少ない語よりも多い語を優先し，第 2 項は T_w の頻度の平均値と定義されている．式 1 は第 2 項を期待値に変更することで，他の語の一部としてよく用いられる語に，厳しいペナルティを与えている．

[9] 略語の短縮形と完全形以外の関係が括弧書きで示される例としては，*serotonin (5-HT)* のような類義語関係がある．

表 3: MEDLINE で頻出する略語（上位 10 件を抜粋）

略語	略語を含む文数	完全形の種類数
CT	30982	171
PCR	25387	39
HIV	19566	13
LPS	18071	51
MRI	16966	18
ELISA	16527	25
SD	15760	165
BP	14860	145
DA	14518	129
CSF	14035	34

とされる短縮形・完全形ペアの数が 317 個と少ない．提案手法は，大規模な入力テキストから略語辞書を作ることを目的としているので，52GB の MEDLINE [10]アブストラクト中で頻出する略語 50 件を選び，その略語の短縮形を含む 253, 237 文から，人手で完全形を抜き出し，計 4, 212 件の短縮形・完全形ペアから成る正解データを作成した．表 3 に，MEDLINE アブストラクト中で頻出する略語上位 10 件を示した．

表 4 は，Schwartz ら [13] の手法（ベースライン・システム），提案手法[11]，その両方を組み合わせた場合[12]に抽出された略語の完全形と，作成した正解データに含まれる略語の完全形の一致度合い[13]を測り，評価したものである．提案手法を単独で使用した場合は，略語の短縮形と頻繁に共起する語が正しい完全形を抽出し，精度が高いが，共起頻度の低い語は抽出できず再現率が低い．ベースライン・システムは，再現率が非常に高いが，大規模なテキストに適用すると，間違った完全形を多く抽出してしまうことが分かる．両者を組み合わせた場合に両者の弱点が克服されることから，提案手法が正しい完全形に対して高いスコアを与えるとともに，文字走査で抽出される冗長な完全形を削除し，精度と適合率のバランスを保っていることが分かる．

4 結論

本稿では，大規模なテキストから略語の短縮形と完全形のペアを自動抽出する手法として，共起情報に基づいて短縮形と完全形のペアのスコア付けを行う手法を提案した．既存の文字走査に基づく手法と組み合わせることで，精度と再現率の両方において約 80% の性能を示した．

[10] `medline05n0001.xml` から `medline05n0500.xml` まで．

[11] $\mathrm{TH}(w) \geq 4$ の語を完全形候補として抽出する．

[12] $\mathrm{TH}(w) < 4$ の語に対して Schwartz らの手法を適用し，完全形候補を補充する．

[13] システムが抽出した完全形と，正解データの完全形を比較する際は，両者にステミングを適用して語形を統一している．

表 4: 略語抽出システムの評価結果

システム	精度	再現率	F スコア
文字走査 [13]	0.555	0.933	0.681
提案手法	0.815	0.140	0.216
提案手法+文字走査	0.787	0.849	0.809

参考文献

[1] E. Adar. SaRAD: A simple and robust abbreviation dictionary. *Bioinformatics*, Vol. 20, No. 4, pp. 527–533, 2004.

[2] S. Ananiadou and G. Nenadic. *Text Mining for Biology and Biomedicine*, chapter Automatic Terminology Management in Biomedicine, pp. 67–97. Artech House, Inc., 2006.

[3] H. Ao and T. Takagi. ALICE: An algorithm to extract abbreviations from MEDLINE. *Journal of the American Medical Informatics Associationj*, Vol. 12, No. 5, pp. 576–586, 2005.

[4] O. Bodenreider. The Unified Medical Language System (UMLS): Integrating biomedical terminology. *Nucleic Acids Research*, Vol. 32, pp. 267–270, 2004.

[5] J. T. Chang and H. Schutze. *Text Mining for Biology and Biomedicine*, chapter Abbreviations in Biomedical Text, pp. 99–119. Artech House, Inc., 2006.

[6] K. Frantzi and S. Ananiadou. The C-value / NC-value domain independent method for multi-word term extraction. *Journal of Natural Language Processing*, Vol. 6, No. 3, pp. 145–179, 1999.

[7] C. Friedman, H. Liu, L. Shagina, S. Johnson, and G. Hripcsak. Evaluating the UMLS as a source of lexical knowledge for medical language processing. In *AMIA Symposium*, pp. 189–193, 2001.

[8] L. S. Larkey, P. Ogilvie, M. A. Price, and B. Tamilio. Acrophile: An automated acronym extractor and server. In *the Fifth ACM International Conference on Digital Libraries*, pp. 205–214, 2000.

[9] G. Nenadic, I. Spasic, and S. Ananiadou. Automatic acronym acquisition and management with domain specific texts. In *LREC-3, 3rd International Conference on Language, Resources and Evaluation*, pp. 2155–2162, 2002.

[10] M. F. Porter. An algorithm for suffix stripping. *Program*, Vol. 14, No. 3, pp. 130–137, 1980.

[11] J. Pustejovsky, J. Castano, B. Cochran, M. Kotecki, and M. Morrell. Automatic extraction of acronym meaning pairs from MEDLINE databases. *MEDINFO 2001*, pp. 371–375, 2001.

[12] J. Pustejovsky, J. Castano, R. Sauri, A. Rumshisky, J. Zhang, and W. Luo. Medstract: creating large-scale information servers for biomedical libraries. In *the ACL-02 Workshop on Natural Language Processing in the libraries*, pp. 85–92, 2002.

[13] A. S. Schwartz and M. A. Hearst. A simple algorithm for identifying abbreviation definitions in biomedical text. In *Pacific Symposium on Biocomputing (PSB 2003)*, No. 8, pp. 451–462, 2003.

[14] J. D. Wren, J. T. Chang, J. Pustejovsky, E. Adar, H. R. Garner, and R. B. Altman. Biomedical term mapping databases. *Database Issue*, Vol. 33, pp. D289–D293, 2005.