大規模コーパスを知識減とした教師無し略語復元

大野正樹 †　　平尾 努 ‡　　永田 昌明 ‡
筧 捷彦 ††
† 早稲田大学基幹理工学研究科
‡ 日本電信電話株式会社 NTT コミュニケーション科学基礎研究所
†† 早稲田大学理工学術院

1　はじめに

本稿では，定義（以下，完全語）を伴わず医療文書中に単独で出現した略語を，文脈に応じてその完全語を正しく復元する手法を提案する．略語は医療文書において頻繁に使用されており，そしてその多くが複数の完全語を持っている [1]．そのため完全語を伴わずに略語が単独で文章に出現した場合，それが指し示す完全語を特定することは非常に困難である．

例えば，下記の 2 つの文[1]を考えよう．

1. The ability of vancomycin to bind three cytoplasmic peptidoglycan precursors in bacterial species was studied using affinity capillary electrophoresis (**ACE**).
2. In this study，we investigated the in vitro **ACE** inhibitory and in vivo antihypertensive effect of insect cell extracts.

これらの文に出現する ACE は，複数の完全語を持つ略語である．(1) の ACE は完全語を伴って出現しているため，この ACE が affinity capillary electrophoresis を示していることが分かる．しかし，(2) の ACE は完全語を伴わず単独で出現しているため，背景知識持った専門家でない限り，その完全語を特定することができず，この文を正しく理解することができない．そのため，背景知識持を持たない人を支援するために，略語が完全語を伴わず単独で出現した場合でも，それが完全語を特定する仕組みが必要である．

略語を正しく復元することは非常に重要であるが，多くの略語は複数の完全語を持っているため，そして対象となる略語は大量にあるため，人手で復元することは困難である．そのため，計算機により自動的に略語の復元をすることが必要である．

[1] (1) は PMID8080114 の一部．(2) は PMID20735247 の一部．ACE は angiotensin converting enzyme を示している．

完全語を伴わずに単独で出現する略語を復元する場合，その略語の完全語候補を把握しなければならない．提案手法は，Gaudan らの手法 [2] を拡張した手法であり，略語・完全語対を用意せずとも，大量のコーパスから自動的に略語・完全語対を作成し，任意の略語の復元を行う．はじめに，略語と完全語が同時に出現する文書から完全語候補 C を獲得し，その後，完全語候補 C をその末尾語に着目してクラスタリングする．さらに，同一クラスタに属する完全語候補に対し共起情報を利用して不要語を削除することで，略語と正確な完全語の対を自動的に構築する．最後に，復元対象の略語を含む文書と各クラスタの完全語の周辺文脈の類似度を求め，最も類似度が高いクラスタに属する完全語を，その略語の完全語とみなす．

MEDLINE に頻出する略語を用いて評価実験を行ったところ，提案手法は従来手法である Gaudan の手法と比較して平均正解率で上回った．

2　関連研究

Pakhomov らは，略語・完全語対を人手で作成し，それを用いて略語を復元した [3]．また，Stevenson らは Schwartz[4] らの手法を利用して，任意の略語の完全語群を獲得し，それらを人手で整理した [1]．Stevenson らの手法と Pakhomov らの手法は，略語復元の難易度の指標となる重要な研究である．しかし，医療文書は大量にあり，これらを人手で全て整理することは難しいため，これらの手法が適用できる範囲は限られる．

Okazaki らは，略語・完全語対を自動で作成し，略語復元に取り組んだ [5]．始めに自身の提案した手法 [6] により，任意の略語の完全語群をコーパスから獲得し，それらを，完全語の距離を素性にした最大エントロピーモデルによりそれらをクラスタリングすることで．同義語をまとめ．略語・完全語対を作成した．

この手法は高精度であるが，学習のためのラベル付きデータが必要であり，そのデータの作成するための人的コストがかかる．本稿では，人的コストを減らすため，ラベル付きデータを用いないことを目指す．

Gauden らも，下記の手続きで略語・完全語対を自動で作成した [2] .

1. Adar らの手法 [7] により，任意の略語 a の完全語候補 $C = \{c_1, c_2, ...c_n\}$ をコーパスから獲得する．
2. 文字 N グラムに基づくコサイン距離により，完全語候補 C をクラスタリングする．
3. クラスタ間の類似度を，完全語候補の周辺文脈のダイス係数で表し，類似度の高いクラスタを併合する．

その後，各クラスタをそのクラスタ内の完全語候補の周辺文脈に基づき，略語 a 含む文書をその文書内容に基づき，それぞれ素性ベクトルに変換する．この際に，C-value[9] を用いて，素性選択を行う．最後に, サポートベクターマシーン（以下，SVM ）により，略語 a を含む文書がどのクラスタに属するか求め，そのクラスタ内に頻出する完全語候補を略語 a の完全語とみなす．

この手法は，人的コストがかからないと言う点で優れた手法と言えるが，復元できる略語に制限がある．Adar らの手法は，人手で定めたルールに基づいて完全語を認識する手法であるが，そこで用いられたルールは完全ではないためである．例えば，Adar らの手法では water activity (AW) などの，略語と異なった並び順で登場する完全語を認識することができない．他にも，人手で定めたルールの問題点として，ルールを作成するために時間がかかりすぎること，時間が経つにつれルールのメンテナンスコストが大きくなることが指摘されている [8] .

3 教師無し略語復元

提案手法は，大規模コーパスを用いて略語・完全語対を自動で作成し，それを用いて略語を復元する．この手法は，略語・完全語対を作成する際にラベル付きデータを用いないため，教師なし手法である．手法の大枠は Gauden 手法と同等であるが，Gauden 手法では獲得できなかった完全語を，提案手法は獲得することができる．

3.1 完全語候補の獲得

略語 a とその完全語候補 c が同時に出現する含む文書群 $D = \{d_1, d_2, ...d_n\}$ から，完全語候補 $C = \{c_1, c_2, ...c_n\}$ を抽出する．文書 D を，以下ではコーパスと呼ぶ．文書から完全語を獲得する際は，下記の仮定を置いた．

- 「完全語 (略語) 」というパターンで略語・完全語対が出現する
- 略語 a の完全語は，a の出現した位置から min(|a|+5, 2*|a|) 語以内に存在する（ |a| は a の文字数を表す）
- 略語に含まれる語は，英数字は完全語の先頭に出現する

これらの仮定に基づいて完全語候補の探索範囲を定めて，探索範囲の両側から単語を検査し，完全語になり得ない単語を探索範囲から除外する．

左側からの探索は，括弧表現から min(|a|+5, 2*|a|) 語前，または括弧表現直前の機能語から始める．略語中の文字を先頭に持つ単語が出現するまで，右側に探索範囲を狭める．右側の探索は，括弧表現の直前の単語から始める．略語中の文字を先頭に持つ単語が出現するまで，左側に探索範囲を狭める．両側からの探索が終了した後，探索範囲に存在する単語を完全語候補とみなす．探索範囲に単語が存在しない場合，文書中に完全語が出現しなかったとみなす．

従来のルールベース手法 [4][7] よりも弱い制約を用いて探索を行うことにより，water activity (AW) などの，語の並び替えが起きている略語・完全語対も獲得することができる．ただし，制約が弱いため，完全語候補 C に不要な語が混じることがある．また，完全語候補 C の中に同義語があるが，語形が異なるため，それらが同義語として認識できない．よって，次に示す手法で，完全語候補のクラスタリングを行い，同義語をまとめ，不要な語を削る．

3.2 完全語候補のクラスタリングとクラスタを代表する完全語の決定

人がある完全語をもとにして新しい略語を作ろうとするとき，

1. 既存の略語に類似しないような略語を考える
2. または，既存の略語に重複しないような完全語を考える

と本稿では考えた．

例として，major capsid protein : MCP が広く知られている状況を考えよう．ここで，新しいたんぱく質 (protein) を発見し，membrane cofactor protein と名づけ，その略称を MCP にしようとしても，先の MCP が広く知れ渡っていることに気づけば，名前（完全語）そのものを変更するか，略称を変更するか，いずれかの手段をとるであろう．

そして，本稿では，完全語の意味的な中心を担っているのは主辞である末尾の単語であるから，それらが異なっていると仮定し，3.1 で獲得した完全語をクラスタリングする．

上記の仮定に従って，完全語候補 C とそれに対応する文書 D をクラスタリングする．完全語候補の末尾の語が，略語中の全ての文字を含んでいるなら末尾の 1 語，そうでないなら末尾の 2 語をクラスタリングのキーとする．

次に，クラスタ内の完全語の出現頻度によって，クラスタの代表語を決定する．具体的には，末尾の単語を頂点，共起した語を節点として，出現頻度の木をつくり，根から葉に向かって探索をすすめる．ある節点が複数の節点を子として持っていた場合，出現頻度が最も高い節点に進む．また，頂点の出現頻度の θ %を閾値とし，出現頻度が閾値未満の節点を探索の対象に含めない．探索が終了したときに，頂点から探索済みの節点をそのクラスタの代表語とみなす．

この手法は，出現頻度が低い完全語を認識することが困難であるという問題を抱えているが，大規模なコーパスを用いることで，その問題を最小限に抑えることが可能であると考えている．

3.3 完全語の決定

略語 a を含む文書 d_a がどのクラスタに属するか文書分類問題を解き，略語 a の完全語を決定する．文書分類問題を解くために，様々な手法があるが，本稿では k 近傍法（以下，k-NN）と SVM を試した．

始めに，コーパス D を文書内容に基づき素性ベクトル $V = \{v_1, v_2, ...v_n\}$ に，文書 d_a をその文書内容に基づき，それぞれを素性ベクトル v_a に変換する．

k-NN を使った手法では，始めに v_a と V の要素との cosine 距離を計算し，距離が近い素性ベクトル $v_i(\in V)$ を k 件選ぶ．そして，選んだ k 件の素性ベクトルが属するクラスタを調べ，最も頻出するクラスタの代表語を，略語 a が指し示す完全語とする．

SVM を使った手法では，各クラスタに属する素性ベクトルを訓練データとして学習し，分類器を作成する．そして分類器によって，v_a がどのクラスタの属するかを判定する最後に，そのクラスタの代表語を，略語 a が指し示す完全語とする．

4 評価実験

4.1 実験概要

提案手法の有効性を確認するために，MEDLINE に頻出する 15 種類の多義性を持つ略語を対象として，評価実験を行った．実験は，テストセットに出現する完全語を伴わずに単独で出現した略語のうち，何件を正しく復元できたか，その正解率を測るものである．

実験に用いる文書群は，MEDLINE アブストラクトから得た．15 種類の略語を含む文書群に対して，下記の手続きで評価用テストセットを作成した．表 1 に，実験に用いたデータを示す．

1. 略語と完全語が同時に出現している文書を人手で検査し，その略語の完全語候補 C を獲得する．
2. 出現回数が 30 回以上の完全語候補を C_{major}，出現回数が 15 回以上 30 回未満の完全語候補を C_{minor} とする．C_{major} に属する語を含む文書を 20 件，C_{minor} に属する語を含む文書を 5 件選びそれらの文書をテストセット T とした．
3. 各略語に対し，テストセットに利用しなかった文書を，略語・完全語対を作成するための訓練データとした．

比較対象として，Gauden らの手法と，人手で作成した辞書を使用し，k-NN によって曖昧性の解消を行う手法を用意した．Gauden らの手法はプログラムが公開されていなかったため，論文 [2] に基づき筆者が実装した．

各々の手法に用いられるパラメタはテストセットの一部を用いて決定した．提案手法では $\theta = 50$ と設定した．また，bi-gram と uni-gram を用いて素性ベクトルを作成し，k=3 とした k-NN と，線形カーネルを使った SVM を使用して文書分類を行った，

4.2 結果と考察

実験結果の平均（マクロ平均）を表 2 に示す．素性やアルゴリズムを変更したときでも，提案手法が Gauden らの手法を平均正解率で上回った．この結果から，提

表 1: 実験に用いたテストセット

a	$\mid C_{major} \mid$	$\mid C_{minor} \mid$	|T|
AMI	4	0	80
BMD	5	1	105
CAD	10	3	215
CHO	4	1	85
CNS	5	0	100
CRP	4	3	95
HBV	1	1	25
LDL	2	0	40
MDA	7	4	160
PBMC	2	0	40
PMA	5	4	120
SEM	3	2	70
SOD	3	0	60
TEM	5	1	105
TPA	4	5	105

案手法の方が Gauden らの手法よりも正確に略語・完全語対を構築できていることが分かる．また完全語の末尾の 1 語または 2 語が異なっているという，クラスタリングの際に用いた仮定は，コーパス内の完全語に対して正しかった．

Gauden らの手法で構築した略語完全語対と，C-value を素性として SVM を用いたときの正解率は 0.235 であった．Gauden らの手法の正解率が低い原因として，素性に C-value を用いていることが挙げられる．素性を本手法と同等なものに変更したところ，正解率が向上した．

どちらの手法も正解率が最も高いのは，uni-gram を素性として SVM を用いたときである．提案手法の正解率が 0.720，Gaudan らの手法の正解率が 0.648 であり，その差は 7.2 ポイントであった．

表 2: 提案手法とベースライン手法の平均正解率

	SVM C-value	SVM uni	SVM bi	k-NN uni	k-NN bi
提案手法	-	**0.720**	0.623	0.317	0.644
Gaudan	0.235	**0.648**	0.562	0.307	0.544
人手辞書	-	0.797	0.682	0.319	0.657

5 おわりに

本稿では，完全語を伴わず単独で出現した略語の完全語を，大規模コーパスを用いて復元する教師無し手法を提案した．提案手法は，略語・完全語対を用意せずとも，自動的に略語・完全語対を作成し，略語を復元を行う．大規模コーパスから自動で略語・完全語対を獲得する場合，正確に完全語を獲得することが重要であり，それを効率良く行うために，文書から獲得した完全語候補の末尾単語に着目したクラスタリング手法，クラスタ内の共起情報を用いた完全語の決定法を新たに提案した．

評価実験を行っ たところ，提案手法は従来手法である Gaudan の手 法と比較して平均正解率で上回った．

参考文献

[1] Stevenson, M., Guo, Y., Gaizauskas, R. and Martinez, D.: Disambiguation of biomedical text using diverse sources of information., *Proceedings of the BioNLP 2009 Workshop* (2009).

[2] Gaudan, S. and Kirsch, H.: Resolving abbreviations to their senses in Medline, *Bioinformatics*, Vol. 21, No. 18, pp. 3658–3664 (2005).

[3] Pakhomov, S., Pedersen, T. and Chute, C. G.: Abbreviation and Acronym Disambiguation in Clinical Discourse, *Proceedings Proceedings of AMIA Symposium(AMIA2005)* (2005).

[4] S.Schwartz, A. and HEARST, M. A.: A Simple Algorithm for Identifying Abbreviation, *Paci c Symposium on Biocomputing(PSB 2003)* (2003).

[5] Okazaki, N., Ananiadou, S. and Tsujii, J.: Building a high-quality sense inventory for improved abbreviation disambiguation, *Bioinformatics*, Vol. 26, No. 9, pp. 1246–1253 (2010).

[6] Okazaki, N. and Ananiadou, S.: Building an abbreviation dictionary using a term recognition approach, *Bioinformatics*, Vol. 22, No. 24, pp. 3089–3095 (2006).

[7] Adar, E.: SaRAD : A Simple and Robust Abbreviation Dictionary, *Bioinformatics*, Vol. 20, No. 4, pp. 527–533 (2004).

[8] Liu, H., Lussier, Y. A. and Friedman, C.: Disambiguating Ambiguous Biomedical Terms in Biomedical Narrative Text : An Unsupervised Method, *Journal of Biomedical Informatics*, Vol. 34, pp. 249–261 (2001).

[9] Frantzi, K. and Ananiadou, S. and Mima, H.:Automatic recognition of multi-word terms:. the C-value/NC-value method, *International Journal on Digital Libraries*, Vol. 3, No. 3, pp. 115–130 (2000).