情報処理学会論文誌　データベース　Vol. 4　No. 3　45–52 (Oct. 2011)

曖昧性を持った医療略語の教師なし復元手法

大野 正樹[†1]　平尾 努[†2]
永田 昌明[†2]　筧 捷彦[†3]

医学生物学文書中に出現する略語の多くは複数の定義を持っており，それを正しく復元することは大きな課題である．本稿では，英文の医学生物学文書中に定義をともなわずに単独で出現した略語を，文脈に応じて正しく復元する手法を提案する．はじめに，医学生物学文書群から，略語と定義候補の対を含む文書群を獲得し，定義候補群と対応する文書群をクラスタリングする．そして，クラスタを代表する定義を決定し，そのクラスタの情報を用いて，略語復元を行う．定義候補群のクラスタリングには，それらの末尾 2 単語に着目したクラスタリング手法，クラスタを代表する定義の決定には単語の共起情報に着目した手法を，それぞれ提案する．過去の論文で用いられた 20 語と，MEDLINE に頻出する 40 語の計 60 語を実験を用いて評価データを作成し，評価実験を行ったところ，復元精度で従来手法を上回り提案手法の有効性を確認できた．

An Unsupervised Method for Abbreviation Expansion in Biomedical Texts

Masaki Ono,[†1] Tsutomu Hirao,[†2] Masaaki Nagata[†2]
and Katsuhiko Kakehi[†3]

Most of abbreviations in the bio-medical domain have several definitions and to know the correct definition of the abbreviation is a big challenge. The aim of this study is to expand an abbreviation into its intended sequence of words in a full automatic way. The present method consists of three primary steps. First, we extract the documents which have a given abbreviation and its expansion from a large-scale corpus by simple rules. Second, we categorize the expansion forms into clusters and determine the representative expansion of each cluster by an unsupervised method. Third, we determine the most suitable expansion for a given abbreviation based on the cluster contexts. In the second step, we newly use the rightmost two words of the expansion for clustering the expansions and the co-occurrence frequency for determining the representative expansion.

1. はじめに

本稿では，英文の医学生物学文書中に単独で出現した略語を，文脈に応じて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定義（以下，完全語と呼ぶ）を持っている[1)]．完全語をともなわずに略語が単独で文書に出現した場合，それが示す完全語を特定することは非常に困難である．

完全語をともなわずに略語が単独で出現する例として，MEDLINE[⋆1]から獲得した次の文[⋆2]を考えよう．

> In this study, we investigated the in vitro **ACE** inhibitory and in vivo antihypertensive effect of insect cell extracts.

この文に出現する “ACE” には，“affinity capillary electrophoresis” や，“angiotensin converting enzyme” などの複数の完全語が存在する．“ACE” の完全語が一意に定まらないため，背景知識を持った専門家でない限り，“ACE” 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．このように，ある略語が複数の完全語を持つことを，略語が曖昧性を持つと呼ぶ．

略語が完全語をともなわず単独で出現した場合でも，その略語の完全語を特定することができれば，テキストマイニング技術の精度を向上させることができる[2)]．そのため，曖昧性を持った略語を正しく復元することは非常に重要である．

医学生物学文書および複数の完全語を持つ略語は大量にあり，これらを人手で整理することが困難である．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．2011 年 1 月現在，MEDLINE には 1,500 万超の文書が保存されており，年間数十万のペースで新たな文書が追加されている．Okazaki らは，UniProt[⋆3]に保存されている 46 万略語

†1 早稲田大学大学院基幹理工学研究科
Graduate School of Fundamental Science and Engineering, Waseda University
†2 NTT コミュニケーション科学基礎研究所
NTT Communication Science Labs
†3 早稲田大学理工学術院
Faculty of Science and Engineering, Waseda University
⋆1 医学生物学文書のデータベース．保存された文書には，それぞれ固有の ID（PMID）が割り振られている．
⋆2 PMID8080114 の一部．ACE は angiotensin converting enzyme を示している．
⋆3 Universal Protein Resource．タンパク質アミノ酸配列などのタンパク質についての情報を保存する医学系データベース．

© 2011 Information Processing Society of Japan

のうち，32%が 2 つ以上の意味を持ち，1%が 30 以上の意味を持つと報告している[3]．Liu らは，MEDLINE に出現する略語は，全体の 81.2%が 2 個以上の意味を持ち，1 つの略語は平均して 16.6 個の意味を持つと報告している[4]．

完全語をともなわずに単独で出現する略語を復元する場合，その略語と完全語候補の対を何らかの形で辞書として整理しておかなければならない．本稿では，略語の完全語候補とその周辺文脈をまとめたものを略語・完全語辞書と呼ぶ．これらを人手で整備することは多大なコストがかかるため，本稿では，医学生物学文書群から自動で略語・完全語辞書を獲得し，それを用いて略語を復元する手法を提案する．

提案手法では，復元の対象となる略語 a とそれを含む文書 d_a が与えられたときに，次の (1) から (3) の手続きでその完全語を復元する．

(1)　略語 a とその完全語候補を含む文書群 $D = \{d_1, d_2, \ldots, d_n\}$ を医学生物学文書群から取得し，完全語候補群 $C = \{c_1, c_2, \ldots, c_n\}$ を抽出する．

(2)　完全語候補群 C をクラスタリングし，同義語をまとめる．同時に，それに対応する文書群 D を完全語候補群 C に応じてクラスタリングする．そして，各クラスタの完全語候補を名寄せして，クラスタを代表する完全語を決定する．

(3)　文書 d_a をその文書内容に基づき，特徴ベクトル v_a に変換する．同様に，文書群 D の各々の要素を，その文書内容に基づき特徴ベクトルに変換し，特徴ベクトル群 $V = \{v_1, v_2, \ldots, v_n\}$ を作成する．そして，サポートベクタマシーン（Support Vector Machine，以下 SVM）を用いて，特徴ベクトル v_a が属するクラスタを決定する．最後に，そのクラスタを代表する完全語を，略語 a が示す完全語とする．

本稿では (2) において，完全語候補の末尾 2 単語に着目したクラスタリング手法とクラスタ内の共起情報に着目した代表語の決定法を新たに提案する．

過去の論文[1] で用いられた 20 語に，MEDLINE に頻出する語の中から選んだものを加えた計 60 語を対象に評価実験を行ったところ，復元精度は 80.2%で，従来手法を 11.4 ポイント上回り，その有効性を確認した．

2. 関 連 研 究

完全語をともなわずに単独で出現する略語を復元する場合，その略語と完全語候補の対を整理しておかなければならない．本稿では略語・完全語辞書の作成者に着目し，略語復元手法を略語・完全語辞書の作成を，人手で行う手法と計算機により自動で行う手法に分類した．

2.1　略語・完全語辞書を人手で作成する手法

Stevenson らは，Schwartz ら[5] の手法により医学生物学文書群から完全語候補群を自動的に抽出し，それらを人手で整理して略語・完全語辞書を作成した[1]．

Pakhomov らは，教師あり手法と半教師あり手法を試した[6]．教師あり手法では，用意した文書群 D から人手で略語・完全語辞書を作成した．半教師あり手法では，人手で作成した略語・完全語辞書をもとに，文書群 D とは異なる文書群を用いて自動で拡張し，略語・完全語辞書を作成した．

Stevenson らの手法と Pakhomov らの手法は，略語復元を，略語の指し示す完全語をその略語の語義と見なし，略語・完全語辞書をあらかじめ用意しておくことで，略語復元を一種の教師あり語義曖昧性解消問題として扱った．これらの研究は，略語復元の難易度の指標となる重要なものであるが，医学生物学文書は大量にあり人手ですべて整理することが難しいため，これらの手法が適用できる範囲は限られる．大量にある医学生物学文書に対応できるように，略語・完全語辞書を作成するために何らかの工夫が必要である．

2.2　略語・完全語辞書を自動で作成する手法

辞書の作成や保守のコストを小さくするために，Okazaki らと Gauden らは，医学生物学文書群から自動的に略語・完全語辞書を獲得した．

Okazaki らは，自身の提案した手法[7] により医学生物学文書群から完全語群を自動的に抽出し，人手で用意したラベル付きデータで作成した分類器を使用して同義語をまとめ，略語・完全語辞書を作成した[3]．この手法は高精度であるが，分類器の学習のためにラベル付きデータが必要であり，そのデータを作成するための人的コストがかかる．

また Gauden らも，次の (1) から (3) までの手続きで略語・完全語辞書を自動で作成した[8]．

(1)　Adar の手法[9] により，復元対象の略語 a の完全語候補 $C = \{c_1, c_2, \ldots, c_n\}$ を医学生物学文書群から獲得する．

(2)　文字 n-gram 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．

この手法は，Okazaki ら[3] の手法とは異なり，訓練データを作るための人的コストがかからない点で優れた手法といえる．しかし，Adar の手法はルールベースの手法であり，そこで用いられたルールは完全ではないため，Adar の手法では獲得できない完全語がある．たとえば，Adar の手法では完全語の文字から略語が生成される過程で，単語の並べ替えが

© 2011 Information Processing Society of Japan

起きることを考慮していない．そのため，完全語と異なった並び順で登場する略語を認識することができない．他にも，人手で調整し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，作成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という問題がある[10]．

3. 提 案 手 法

本稿では，Gauden らの手法の自然な拡張として，略語・完全語辞書を自動で作成し略語復元を行う手法を提案する．教師なし手法で略語の復元を行う場合，クラスタリングの精度とクラスタの代表語の決定が重要である．本稿では，クラスタリングの手法と，クラスタの代表となる完全語の決定手法を新たに提案する．提案手法は，略語 a とそれを含む文書 d_a が与えられたときに，3.1 から 3.3 節までの手続きで，a の完全語を復元する．

3.1 完全語候補の獲得

略語 a とその完全語候補を含む文書群 $D = \{d_1, d_2, \ldots, d_n\}$ を医学生物学文書群から取得し，完全語候補群 $C = \{c_1, c_2, \ldots, c_n\}$ を抽出する．文書から完全語を獲得する際は，次の仮定をおく．

- 略語 a の完全語は，a の出現した位置から $min(|a|+5, 2\cdot|a|)$ 語以内に存在する[11]（$|a|$ は a の文字数を表す）．
- 略語中のすべての英数字は完全語に出現する．
- 完全語から略語が生成されるときに，文字の並び順が変わることがある．

これらの仮定に基づき，完全語候補の探索範囲を決め，探索範囲の両側から単語を検査し，完全語になりえ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 $min(|a|+5, 2\cdot|a|)$ 語前，または括弧表現直前の機能語から始める．略語の先頭または先頭から 2 番目の英数字を文字列中のどこかに含む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．略語の末尾または末尾から 2 番目の英数字を文字列中のどこかに含む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見なす．探索範囲に単語が存在しない場合，探索を失敗したと見なす．

従来のルールベース手法[5),9)] 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，語の並べ替えが起きている略語・完全語辞書も獲得することが可能である．

しかし制約が弱いため，完全語候補群 C に不要な語が混じることがある．また，完全語候補群 C の中に同義語であっても語形の異なるものが含まれることもある．こうした同義語の例を**表 1** に示す．

表 1　同義語の例
Table 1　Examples of the synonyms.

例 1	exponentially modified gaussian exponential modified gaussian
例 2	mono calcium phosphate monobasic calcium phosphate monocalcium phosphate
例 3	child's apperception test children's apperception test child apperception test children apperception test
例 4	high performance liquid chromatography high pressure liquid chromatography high power liquid chromatography

これらを解決するために，3.2 節に示す手法で，完全語候補群のクラスタリングを行い同義語をまとめ，各クラスタの代表となる完全語を決定する．

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

本稿では，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，次のいずれかの方法をとると考えた．

- 既存の略語に類似しないように略語を選ぶ．
- 既存の略語に重複しない略語となるように，完全語そのものを選ぶ．

さらに，完全語の意味の中心的な役割を担う語は末尾の単語であると考えた．そのため，本稿では，「ある略語が完全語を複数持っている場合，それらの完全語の末尾の 1 語または 2 語は互いに異なっている」という仮定をおいた．末尾の 1 語または 2 語としたのは，末尾 1 単語だけでは，意味をなさない場合があるためである．

例として，angiotensin converting enzyme: ACE が広く知られている状況を考えよう．ここで，新しい酵素（enzyme）を発見し，anonymous converting enzyme と名づけ，その略称を ACE にしようとしても，先の ACE が広く知れわた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．

上記の仮定に従い，完全語候補群 C を同義語をまとめる目的でクラスタリングを行う．そのキーとしては，完全語候補の末尾の語が略語のすべての文字を含んでいるなら末尾の 1 語

© 2011 Information Processing Society of Japan

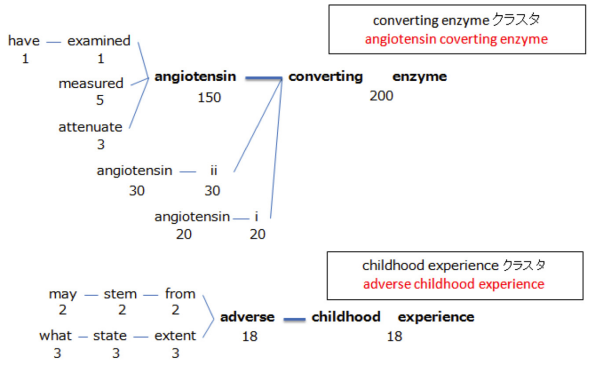

図 1　末尾の 2 単語に着目したクラスタリング手法
Fig. 1　Clustering method based on the rightmost word.

をあて，そうでなければ末尾の 2 語をあてる．この結果に応じて，文書群 D もクラスタリングする．

続いて，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．出現頻度を効率良く計算するために，トライ型の木構造である，頻出パターン木（frequent pattern tree，あるいは FP-tree）[12] を用いる．具体的には，完全語の末尾の単語を頂点とし，共起した語を節点として，出現頻度の情報を持った木をつくる．

各クラスタを代表する語を探すために，根から葉に向かって探索を進める．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，出現頻度が頂点の出現頻度の θ%未満の節点は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見なす．

図 1 はクラスタリングとクラスタを代表する完全語の決定過程を示している．末尾の単語は enzyme であり，略語中のすべての文字を含んでいないため，converting enzyme をキーにする．

この手法は，語の並べ替えが起きている完全語も獲得することができるが，共起情報を利用しているために，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かかえている．本稿では，大規模な文書群を完全語の獲得の対象にすることにより，こ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　略語・完全語辞書を用いた略語復元

略語と完全語の関係が分かれば，略語の周辺文脈から，完全語を推定することが可能となる．これは，多値の文書分類として定式化することができる．ここでは略語 a を含む文書 d_a がどのクラスタに属するか決定する分類問題を解くことで，略語 a の示す完全語を決定する．

文書 d_a をその文書内容に基づき，特徴ベクトル v_a に変換する．同様に，文書群 D の各々の要素を，その文書内容に基づき特徴ベクトルに変換し，特徴ベクトル群 $V = \{v_1, v_2, \ldots, v_n\}$ を作成する．そして，SVM を用いて，特徴ベクトル v_a と類似したクラスタを決定する．最後に，そのクラスタを代表する完全語を，略語 a が示す完全語とする．

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評価実験を行った．実験では，作成した略語・完全語辞書の正確さと，略語復元の正解率の 2 つの観点で評価を行った．前者は，医学生物学文書群から自動で作成された略語・完全語辞書の適合率（precision），再現率（recall），F 値（F-measure）で評価する．後者は，完全語をともなわずに単独で出現した略語とそれを含む文書群を評価データとし，それが含む略語のうち何件を正しく復元できたか，その正解率で評価する．

実験に用いる略語には，過去の論文[1] で用いられた 20 語に，MEDLINE に頻出する語の中から選んだものを加えて，計 60 語を選んだ．先行研究はいずれも 3 文字略語だけを実験対象としてきたが，この実験では，2 文字略語 10 語を含めて対象略語を選び，2 文字略語に対する有効性も調べた．4 文字以上の略語は，完全語数が少ないことから先行研究にならって実験対象としなかった．

提案手法の比較対象として，Gauden らの手法と，人手で作成した略語・完全語辞書を利用した手法を用意した．Gauden らの手法はプログラムが公開されていなかったため，論文に基づき筆者が実装した．

各々の手法に用いるパラメータは，開発データを用いて決定した．開発データは，MEDLINE に頻出する 10 語の略語とした．これにより，提案手法でクラスタの代表語を決定するために使用するパラメータ θ は 50 に設定した．

また，SVM を用いて文書分類を行う際には，unigram，bigram，そして unigram と bigram の組合せを特徴にして，bag-of-words モデルを用いて文書を特徴ベクトルに変換した．この際に，文書中の語に対してステミング処理や小文字化などは行っていない．

4.2　実験手順

実験に用いる文書群は，MEDLINE から得た．60 語の対象略語それぞれに対して，それ

© 2011 Information Processing Society of Japan

を含む文書群 D を MEDLINE から抽出し，次の (1) から (3) までの手続きで，評価データと略語・完全語辞書を作成するための医学生物学文書群を作成した．

(1)　文書群 D を人手で検査し，その略語の完全語候補群 C を獲得する．そして，人手で完全語候補群 C をクラスタリングし，同義語をまとめる．

(2)　各クラスタに対して (a) または (b) の手続きを行い文書を集め，それらを評価データとする．

(a)　クラスタ内の語の総出現回数が 30 回以上なら，そのいずれかを含む文書を 20 件集める．

(b)　クラスタ内の語の総出現回数が 15 回以上 30 回未満なら，そのいずれかを含む文書を 5 件集める．

(3)　文書群 D のうち，評価データに用いなかった文書群を略語・完全語辞書を作成するための医学生物学文書群とする．

60 語の対象略語についてみると，各々の略語が平均して，語の総出現回数が 30 回以上のクラスタを 8.90 個，15 回以上 30 回未満のクラスタを 4.53 個持っていた．

本稿の目的は略語が完全語をともなわずに単独で出現した際に，文脈に応じてその完全語を復元することにある．このため，本来であれば，略語が単独で出現した文書を抽出し，その完全語を人手で与えたデータを評価データとして用意すべきであるが，これには人的コストが多大にかかる．よって，略語・完全語の双方が出現する文書に対し，完全語を削除することで，擬似的に略語が単独で出現したものと見なし，評価データとした．

なお，システムが復元した完全語が削除された完全語と完全一致しなくても，同義語であれば正解とした．

4.3　結果と考察

4.3.1　自動で作成された略語・完全語辞書の結果

自動で作成した略語・完全語辞書の適合率，再現率，F 値を**表 2** に示す．

F 値の平均（マクロ平均）は，提案手法が 0.876，Gaudan 手法が 0.772 であった．そして，F 値の標準偏差は，提案手法が 0.108，Gaudan 手法が 0.141 であった．これらより，提案手法は Gaudan 手法と比較して F 値の平均が高く，ばらつきも小さいことから，その有効性が示せた．

また，提案手法は，2 文字略語・3 文字略語ともに F 値が高い．この結果から，提案手法が 2 文字略語・3 文字略語ともに対して有効であることが分かる．

図 2 に，対象略語の各々に対して自動で作成した略語・完全語辞書の F 値を示す．提案手法は，図 2 でも多くの略語で Gaudan 手法を上回っており，有効性が高いことが分かる．

表 2　自動で作成された略語・完全語辞書の適合率，再現率，F 値
Table 2　Evaluation results (Precision, Recall, F-measure).

		提案手法			Gaudan 手法		
		適合率	再現率	F 値	適合率	再現率	F 値
全体	平均	0.836	0.939	0.876	0.838	0.763	0.772
	標準偏差	0.152	0.094	0.108	0.182	0.189	0.141
2 文字略語	平均	0.931	0.950	0.939	0.839	0.539	0.629
	標準偏差	0.034	0.049	0.031	0.224	0.080	0.114
3 文字略語	平均	0.817	0.937	0.863	0.838	0.808	0.800
	標準偏差	0.159	0.101	0.114	0.175	0.172	0.129

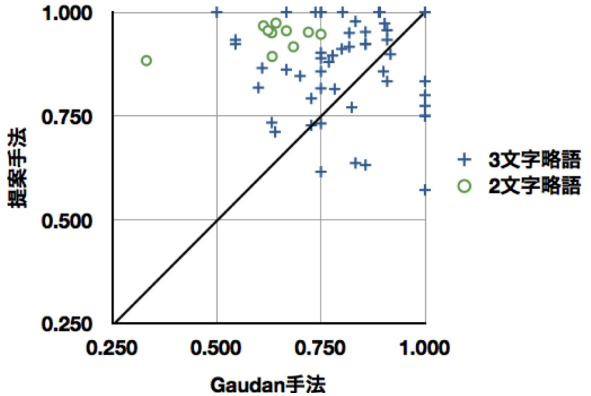

図 2　提案手法と Gaudan 手法の F 値の比較
Fig. 2　Comparing F-measure of proposed method and Gaudan's method.

ある略語に対する完全語候補を集めたものを末尾の 1 語または 2 語で分類してみると，同じグループに分類された語はいずれも同義語であった．そのため，「ある略語が完全語を複数持っている場合，それらの完全語の末尾の 1 語または 2 語は互いに異なっている」という仮定に基づくクラスタリングは今回の実験データにおいては正しく機能した．

F 値の平均では提案手法が Gaudan 手法を上回ったが，個々の結果を見てみると，Gaudan 手法が提案手法を上回った場合がある．提案手法と Gaudan 手法それぞれについて，完全語が獲得できなかったのがどのような場合であるのかを調べた．

提案手法において完全語の獲得が失敗した理由として，完全語として不要な語を削除でき

© 2011 Information Processing Society of Japan

なかったことがあげられる．提案手法は，従来のルールベース手法よりも弱い制約を用いて完全語を獲得し，末尾 2 語との共起情報を用いて不要な語を削除する．このため，そもそも完全語の出現回数そのものが少ない場合には，不要な語が削除されず完全語候補に残る傾向にある．この欠点は，略語・完全語辞書を作成するための医学生物学文書を増やすことで解消できると考える．

Gaudan 手法において完全語の獲得が失敗した理由を 2 つあげられる．1 つ目の理由として，Gaudan が完全語の獲得に用いた Adar の手法が不完全だったことがあげられる．Adar の手法はルールベースの手法であり，そこで用いられたルールが完全語を獲得するには不十分であった．そのため，完全語ではない語を文書から獲得する場合があった．

2 つ目の理由として，完全語のクラスタリングに失敗したことがあげられる．Gaudan 手法では，周辺文脈の類似度がある閾値以上の完全語候補をまとめるが，今回の実験では，閾値の最適値が安定しなかった．よって，本来まとめるべきものが分離したり，その逆になったりする場合があった．このように Gaudan 手法は，完全語の獲得に失敗する場合が多く，提案手法と再現率に差が出たのではないかと考える．

提案手法のパラメータである θ を変化させたときの F 値を**図 3** に示した．図 3 から，θ が極端に大きいときや小さいときに，F 値が小さいことが分かる．また，$\theta = 50$ が 25〜75 付近では F 値が 0.8 を超えており，安定して高い．実験では開発データを用いてパラメータを調整して $\theta = 50$ としたが，実際には $\theta = 45$ とした方が F 値が高かった．それ以外の

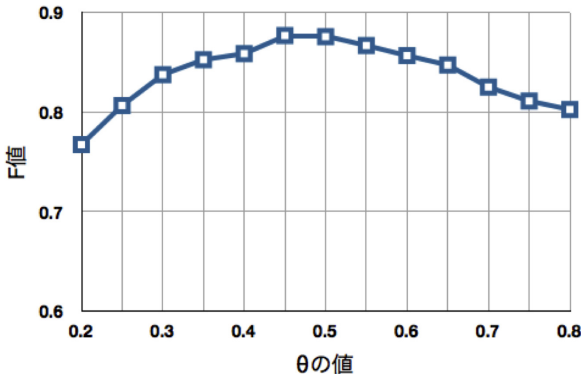

図 3　閾値 θ を変化させたときの F 値
Fig. 3　F-measures on various θ.

θ でも高い F 値を保っていることから，それにさほど敏感ではない．

4.3.2　略語復元の結果

略語復元の正解率を**表 3** に示す．unigram を特徴とした場合を u，bigram を特徴とした場合を b，unigram と bigram を特徴とした場合を u+b と表記した．正解率の平均（マクロ平均）は，unigram を特徴とした場合が最も高く，提案手法が 0.802，Gaudan 手法が 0.688 であった．また，人手で作成した略語・完全語辞書を用いて略語復元を行った場合の正解率が 0.837 であった．この値と提案手法の正解率の割合を求めると，0.958 だった．提案手法は，医学生物学文書群から自動で略語・完全語辞書を獲得し，それを用いて略語を復元するが，その正解率が，人手で略語・完全語辞書を作成した場合の正解率に非常に近いことが分かる．ただし，人手辞書を用いた場合であっても正解率が 0.837 であることから，より高度な分類手法を考えることが今後の課題である．

図 4 に，unigram を特徴とした場合の対象略語各々の正解率を示す．図 4 から，提案手法が Gaudan 手法より正解率が高いケースが多いことが分かる．

どの手法においても，unigram を特徴とした場合に正解率が最も高く，bigram を特徴にした場合に正解率が最も低かった．bigram を特徴にするとデータ数と比較して特徴数が膨大かつ疎になるため，学習がうまくいかなかったのではないかと考える．

4.3.2 項と本項から，略語・完全語辞書の作成と略語の復元の両方の観点で，提案手法が Gaudan らの手法より優れていることを示した．また，クラスタリングの際に用いた，「ある略語が完全語を複数持っている場合，それらの完全語の末尾の 1 語または 2 語は互いに異なっている」という仮定は，今回の評価データにおいて正しかった．そのため，本稿で提案した，完全語候補の末尾単語に着目したクラスタリングと，クラスタ内の共起情報に着目した代表語の決定は有用であると考える．

表 3　略語復元の正解率
Table 3　Accuracy of three methods.

		人手辞書			提案手法			Gaudan 手法		
		u	b	u+b	u	b	u+b	u	b	u+b
全体	平均	0.837	0.748	0.826	0.802	0.719	0.792	0.688	0.630	0.683
	標準偏差	0.129	0.169	0.135	0.146	0.181	0.154	0.216	0.228	0.220
2 文字略語	平均	0.637	0.475	0.614	0.626	0.469	0.599	0.407	0.316	0.388
	標準偏差	0.028	0.053	0.034	0.046	0.058	0.049	0.063	0.064	0.063
3 文字略語	平均	0.874	0.800	0.867	0.835	0.767	0.828	0.741	0.689	0.737
	標準偏差	0.105	0.131	0.108	0.136	0.157	0.142	0.195	0.200	0.196

© 2011 Information Processing Society of Japan

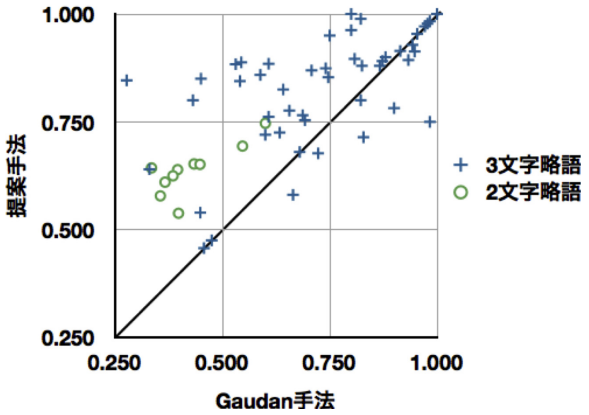

図 4　提案手法と Gaudan 手法の正解率の比較
Fig. 4　Comparing accuracy of proposed method and Gaudan's method.

5. おわりに

本稿では，自動的に略語・完全語辞書を獲得し，それに基づき，単独で出現した略語を文脈に応じて正しく復元する手法を提案した．そして，略語・完全語辞書を作成する際に使用するクラスタリングの手法と，クラスタの代表語の決定手法を新たに提案した．提案手法はラベル付きデータを必要とせず，人的コストが低い．

提案手法では，「ある略語が完全語を複数持っている場合，それらの完全語の末尾の 1 語または 2 語は互いに異なっている」という仮定をおき，獲得した完全語候補群を，その末尾 2 単語に着目してクラスタリングした．そして，クラスタを代表する定義の決定には単語の共起情報に着目した．

過去の論文で用いられた 20 語と，MEDLINE に頻出する 40 語の計 60 語を用いて評価実験を行ったところ，復元精度は 80.2%で，従来手法を 11.4 ポイント上回り，その有効性を確認した．

本稿で提案した手法は人手でラベルを付与したデータを必要としないという特徴がある．そのため，大量に生み出される医療文書に出現する略語を復元する際により有効な方法であると考える．今後は実験対象の略語数を増やし，提案手法で用いた仮定が正しいか検証したい．

謝辞　有益なコメントを頂戴した匿名の査読者に感謝いたします．

参 考 文 献

1) Stevenson, M., Guo, Y., Gaizauskas, R. and Martinez, D.: Disambiguation of biomedical text using diverse sources of information, *Proc. BioNLP 2009 Workshop* (2009).
2) Cohen, A.M. and Hersh, W.R.: A survey of current work in biomedical text mining, *Briefings in Bioinformatics*, Vol.6, No.1, p.57 (2005).
3) 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).
4) Liu, H., Aronson, A.R. and Friedman, C.: A Study of Abbreviations in MEDLINE Abstracts, *Proc. AMIA Symposium* (*AMIA2002*), pp.464–468 (2002).
5) Schwartz, A.S. and Hearst, M.A.: A Simple Algorithm for Identifying Abbreviation, *Pacific Symposium on Biocomputing* (*PSB 2003*) (2003).
6) Pakhomov, S., Pedersen, T. and Chute, C.G.: Abbreviation and Acronym Disambiguation in Clinical Discourse, *Proc. AMIA Symposium* (*AMIA2005*) (2005).
7) Okazaki, N. and Ananiadou, S.: Building an abbreviation dictionary using a term recognition approach, *Bioinformatics*, Vol.22, No.24, pp.3089–3095 (2006).
8) Gaudan, S. and Kirsch, H.: Resolving abbreviations to their senses in Medline, *Bioinformatics*, Vol.21, No.18, pp.3658–3664 (2005).
9) Adar, E.: SaRAD: A Simple and Robust Abbreviation Dictionary, *Bioinformatics*, Vol.20, No.4, pp.527–533 (2004).
10) 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).
11) Park, Y. and Byrd, R.J.: Hybrid text mining for finding abbreviations and their definitions, *2001 Conference on Empirical Methods in Natural Language Processing* (*EMNLP2001*), pp.126–133 (2001).
12) 宇野毅明，有村博紀：頻出パターン発見アルゴリズム入門—アイテム集合からグラフまで，第 22 回人工知能学会全国大会予稿集 (2008).

(平成 23 年 3 月 20 日受付)
(平成 23 年 5 月 2 日採録)

(担当編集委員　村田 真樹)

© 2011 Information Processing Society of Japan

大野　正樹（正会員）

2009 年早稲田大学理工学部コンピュータ・ネットワーク工学科卒業．2011 年早稲田大学大学院基幹理工学研究科修士課程修了．言語処理学会会員．

平尾　　努（正会員）

1995 年関西大学工学部電気工学科卒業．1997 年奈良先端科学技術大学院大学情報科学研究科博士前期課程修了．同年（株）NTT データ入社．2000 年より，NTT コミュニケーション科学基礎研究所に所属．博士（工学）．自然言語処理の研究に従事．言語処理学会，ACL 各会員．

永田　昌明（正会員）

1987 年京都大学大学院工学研究科修士課程修了．工学博士．同年 NTT 入社．1989 年から 4 年間 ATR 自動翻訳電話研究所へ出向．1999 年から 1 年間 AT&T 研究所客員研究員．統計的自然言語処理の研究に従事．現在，NTT コミュニケーション科学基礎研究所主幹研究員．情報処理学会奨励賞（1991 年），情報処理学会論文賞（1995 年），人工知能学会研究奨励賞（1995 年）等受賞．電子情報通信学会，人工知能学会，言語処理学会，ACL 各会員．

筧　　捷彦（フェロー）

早稲田大学理工学術院教授（基幹理工学部情報理工学科）．1970 年東京大学大学院工学系研究科修士課程修了．東京大学工学部助手，立教大学理学部講師・助教授を経て，1986 年早稲田大学理工学部教授，2007 年から現職．日本ソフトウェア科学会，ACM 各会員．情報処理教育委員長．

© 2011 Information Processing Society of Japan