論　文

多様な遺伝子名認識と文書分類を用いたGene Ontologyアノテーション

関　和広[†a)]　　モスタファ ジャビド[††]

Gene Ontology Annotation in a Text Categorization Framework combined with Robust Gene Name Recognition

Kazuhiro SEKI[†a)] and Javed MOSTAFA[††]

あらまし　Gene Ontology（GO）は遺伝子の機能を記述するために作られた一種の統制語彙であり，複数の異なるモデル生物データベースを横断的に検索可能にする．GO コードを個々の遺伝子に付与する作業，すなわち GO アノテーションは情報アクセスの効率性を向上させるために重要であるものの，その専門性の高さから自動化は進んでいない．そこで本研究では GO アノテーション自動化の第一歩として，テキスト分類の手法を応用して GO ドメインの推定を行う．まず，生物医学分野の特殊な術語を考慮し，辞書や規則を併用して多様な遺伝子名表記を同定する．そして同定した遺伝子名出現箇所の周辺単語を当該遺伝子に関連付け，教師付きの単語重みと k 近傍法を用いて適切な GO ドメインを推定する．実データを用いた評価実験により提案手法の優位性を示し，さらに同手法が他の関連分類問題においても有効であることを示す．

キーワード　オントロジー，分類，遺伝子機能，遺伝子名同定，教師付き単語重み付け

1. まえがき

ヒトゲノム計画の完了以降，分子生物学の重要な課題の一つとして，個々の遺伝子の機能同定に関する研究が活発に行われている．マイクロアレイなど高速な遺伝子解析技術の登場にも後押しされ，生物医学分野の学術論文は近年ますます増大している．しかしながら，これら大量の論文は自然言語で記述されているため，所望の情報を網羅的に収集・利用することは容易ではない．これらの文章中に埋もれた有用な情報を整理・構造化し，効率的なアクセスを可能にするため，現在，多くの努力がなされている．その一つが Gene Ontology（GO）によるアノテーションである．GO は様々なモデル生物データベース，例えば，FlyBase，Saccharomyces Genome Database，Mouse Genome Informatics Database などを同一の語彙によって検索可能にするために構築された一種の統制語彙であり，図 1 のような階層構造を持つ．階層構造のルート下の第一層は，三つのドメイン，すなわち molecular function (MF)，cellular component (CC)，biological process (BP) から成る．

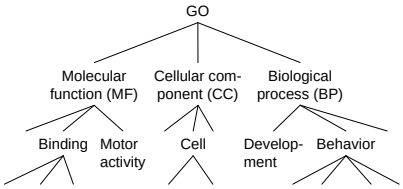

図 1　Gene Ontology の基本構造
Fig. 1　Structure of Gene Ontology.

アノテーションの対象となる遺伝子数・文献数の膨大さと内容の専門性の高さから，GO アノテーションは大変な労力に加えて分野固有の広範な知識を必要とし，curator と呼ばれる専門家によって手作業で行われている．よって，GO アノテーションを自動化あるいは半自動化できれば，人間の負担を大きく軽減できる．このような背景から，GO アノテーションの自動化に関する研究が Text REtrieval Conference (TREC) 2004 の Genomics トラック [1] で行われた．

† 神戸大学自然科学系先端融合研究環，兵庫県
Organization of Advanced Science and Technology
Kobe University
1–1 Rokkodai, Nada, Kobe 657–8501, Japan
†† インディアナ大学，米国
Indiana University
1320 E. 10th St., LI 011, Bloomington, IN 47405, USA
a) E-mail: seki@cs.kobe-u.ac.jp

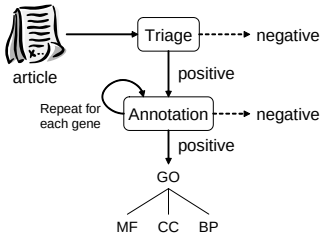

図 2　分類タスクの流れ
Fig. 2　A conceptual flow of the categorization task.

Genomics トラックは生物医学分野のテキストを対象に 2003 年から 5 年間の計画で進行中であり，2004 年度には二つのタスクが設定された．一つは情報検索タスクであり，もう一方が本論文の主題である分類タスクである．分類タスクでは，前述の GO アノテーションを実現するための第一歩として，二つのサブタスク，すなわち Triage タスクと Annotation タスクが設けられた．両者は現在 Mouse Genome Informatics（MGI）で行われている GO アノテーションのプロセスの一部を擬したものである．図 2 にその流れを示す．

Triage タスクは，入力となる論文（article）が GO コードを付与すべき実験的記述を含んでいるかどうかを判断するものであり，具体的な GO コードについては関知しない．Annotation タスクは，Triage タスクに続く処理であり，論文中で言及されている個々の遺伝子について，論文中の記述に応じて適切な GO ドメインの付与を行う．いずれの GO ドメインも不適切な場合は，NEG（negative）と判断する．一件の論文が複数の遺伝子について記述している場合は，それぞれの遺伝子に対して繰り返し GO ドメイン付与を行う．

本研究では，特に Annotation タスクに焦点を当て，テキスト分類の手法を用いて高精度な GO ドメインの推定を目指す．以下，2. で提案する GO ドメイン推定の枠組みについて詳説し，3. で評価実験の概要を述べる．4. で実験結果の結果を報告し，考察を加える．5. で関連研究について議論し，6. で本論文のまとめと今後の課題について述べる．

2. 提 案 手 法

本節では，提案する GO ドメイン推定の枠組みについて，文書表現と分類器の順に詳説する．

2.1 文 書 表 現

2.1.1 テキスト断片の抽出

GO ドメイン推定は個々の論文ではなく，論文中で言及される個々の遺伝子についてその記述内容に応じて行わなければならない．よって，たとえ複数の遺伝子（gene）が一つの論文（article）中で言及されていたとしても，それぞれの $\langle \text{article}, \text{gene} \rangle$ ペアについて個別にドメインの推定を行う必要がある．したがって，それぞれの $\langle \text{article}, \text{gene} \rangle$ ペアをテキスト分類における「テキスト」あるいは「文書」として扱う必要がある．このため本研究では，まずアノテーションの対象となっている遺伝子について言及している箇所周辺だけを同定・抽出し，それらテキスト断片の集合を $\langle \text{article}, \text{gene} \rangle$ に関連付ける．この処理は，遺伝子名拡張と遺伝子名同定に分けることができる．以下，それぞれの処理について述べる．

a） 遺伝子名拡張

遺伝子名拡張は，所与の遺伝子名をその同義語と関連付ける処理である．遺伝子名（またはタンパク質名）は多くの別名，省略形，シンボル名を持つことで知られており，例えば *membrane associated transporter protein* は，*underwhite*, *Matp*, *uw*, *Dbr*, *bls*, *Aim1* などとも呼ばれる．よって，このタンパク質に関する記述を探す場合は，これらすべての同義語を考慮する必要がある．これに対処するため，本研究では入力論文と遺伝子データベースの二種類の情報源を用いた．

まず入力論文に関しては，`<KEYWORD>`と`<GLOSSARY>`フィールドを同義語の情報源として用いた．入力論文は SGML タグで記述されており，この二つのフィールドには遺伝子名とその同義語が明示的に示されていることがある．なお，遺伝子名の同義語（特に省略形）は，論文中で遺伝子名に続いて括弧書きされる場合もある [2]．しかし，括弧表現からの同義語獲得は予備実験で良い結果が得られなかったため，本研究では用いていない．

もう一方の情報源として，遺伝子データベースから遺伝子名とその同義語を自動的に抽出した．本研究では実験的に SWISS-PROT と LocusLink を用いた．この結果，493,473 レコードから成る遺伝子名辞書が構築できた．各レコードは遺伝子あるいはタンパク質の正式名をエントリとして持ち，その（複数の）同義語をリストとして持つ．以降，遺伝子正式名と同義語をまとめて単に「遺伝子名」と呼ぶことにする．

なお，上記のように自動生成した遺伝子名辞書は，同名の遺伝子名やデータベース内の不整合なフォーマットなどに起因して，あまり正確ではないことが報告されている [3]．よって，一般的な用途に用いる遺伝

子名認識システムでは，手作業による修正によって辞書の質を向上させることが重要になる．しかしながら，本研究の対象である GO ドメイン推定に関しては，このような辞書の品質から受ける影響は少ないと考えられる．これは，入力遺伝子に関して辞書が誤った同義語を提示したとしても，そのような無関係な語は入力遺伝子と関連付けられている論文中には現れにくいと考えられるからである．

b） 遺伝子名同定

遺伝子名拡張によって得られた全ての同義遺伝子名に基づき，アノテーションの対象である遺伝子名の出現箇所を同定する．ここで問題となるのは，遺伝子名には記号や空白の恣意的な利用，大・小文字の別などによる異表記が大量に存在することである．これらの微妙な違いを吸収するために，辞書のエントリと入力論文の両方を次の規則によってあらかじめ正規化する．

- アルファベットと数字以外の記号をすべて空白に変換（例：`NF-kappa` → `NF kappa`）
- 異なる文字タイプ（アルファベットと数字）間に空白を挿入（例；`Diet1` → `Diet 1`）
- ギリシャ文字と他の語間に空白を挿入（例：`kappaB` → `kappa B`）
- 全ての英文字を小文字に変換

続いて，遺伝子名の出現箇所を同定するため入力論文を走査する．この際の問題点は，語順・語形の変化や余分な語の挿入などにより，同一であるべき遺伝子名が必ずしも一致しない場合があることである．例えば，*peroxisome proliferator activated receptor binding protein* は *peroxisome proliferator activator receptor (PPAR)-binding protein* のように表記されることがある（下線部が表記の違いを示す）．これに対処するため，本研究では次のように曖昧単語列一致によって遺伝子名を同定する．まず，探索している遺伝子名 *gene* を構成するいずれかの単語と一致する箇所を遺伝子名候補 *candidate* とする．それぞれの候補に関して，次の Overlap スコアを算出する．

$$\mathrm{Overlap}(gene, candidate) = \frac{M - \alpha \cdot U}{N + \beta} \qquad (1)$$

ここで M と U は一致した単語数と一致しなかった単語数を示す．また，α は一致しなかった語のペナルティ（0.3 に設定），N は遺伝子名を構成する単語数，β は短い遺伝子名の Overlap スコアを抑えるための定数（2 に設定）である．もし，ある候補の Overlap スコアが閾値（0.3 に設定）を超えたら，その候補を含む段落を当該遺伝子に関する記述と判断して抽出する．前述の *peroxisome* ... の場合，一致する語が 5 語，一致しない語が 2 語なので，Overlap スコアは 0.55 となる．これは閾値よりも大きいので，この候補を含む段落が対応する $\langle \mathrm{gene}, \mathrm{candidate} \rangle$ ペアと対応付けられる．抽出した段落には PubMed 禁止語リストに基づく禁止語除去，Lovins stemmer [4] による接辞除去を施す．なお，上記のパラメータは，3.1 で述べる訓練データを用いた予備実験によって経験的に設定した．

2.1.2 MeSH 索引語

入力論文から抽出した内容語に加え，本研究では外部の知識資源として Medical Subject Heading（MeSH）を用いる．具体的には，それぞれの入力論文に付与されている全ての MeSH 索引語を MEDLINE データベースから取り出し，これらを当該論文とそこで言及される遺伝子の組に関連付ける．本来は個々の遺伝子に対して関連する MeSH 索引語だけを関連付けることが望ましいが，MeSH 索引語は（遺伝子ではなく）論文に付与されているため，本研究では遺伝子間の区別はしていない．

2.1.3 素 性 選 択

素性選択は，分類を行う上で重要な素性だけを何らかの統計量によって選択する処理である．これによって処理が高速化するだけでなく，一般的に分類精度も向上することが知られている． 本研究では，クラス間で最大のカイ二乗統計値に基づいて素性を選択する $\max \chi^2$ 法 [5] を用いる．なお，素性（単語）の数は予備実験から 3,000 とした．

2.1.4 素性の重み付け

これまでの処理によって，それぞれの入力 $\langle$article, gene$\rangle$ は語の集合で表わされている．次節で述べる kNN を適用するため，これを単語ベクトルへ変換し，TFIDF 法によって重み付けを行う．文書 d における単語 t の TFIDF 値は次式で定義される．なお，ここで文書 d とは $\langle$article, gene$\rangle$ に関連付けられた語の集合を指す．

$$\mathrm{TFIDF}(t,d) = (1 + \log \mathrm{TF}(t,d)) \cdot \log \frac{N}{\mathrm{DF}(t)} \qquad (2)$$

$\mathrm{TF}(t,d)$ は文書 d における語 t の頻度，N は総文書数，$\mathrm{DF}(t)$ は t を含む文書数である．

上記の TFIDF の他に，本研究では，Debole と Sebastiani [6] が提案した教師付き単語重み付け（supervised term weighting）を用いる．教師付き単語重み付けは訓練データ中のラベル付き事例を考慮した方法

であり，素性選択で算出した統計値を IDF の代わりに用いる．ここでは，実験的に次の二種類の重み付けを実装した．$\chi^2(t)$ は語 t のカイ二乗値を表す．

$$\begin{aligned}\mathrm{TFCHI}_1(t,d) &= (1+\log \mathrm{TF}(t,d))\cdot\chi^2(t)\\ \mathrm{TFCHI}_2(t,d) &= (1+\log \mathrm{TF}(t,d))\cdot\log\chi^2(t)\end{aligned} \quad (3)$$

2.2 分 類 器

本研究では k 近傍分類法（kNN）を用いて GO ドメインの付与を試みる．kNN は事例ベースの分類器であり，新聞記事と医学分野のテキスト分類において良好な性能を示すことが報告されている [7]．入力 v が与えられたとき，kNN は v の近傍の k 事例がどのクラスに属するかによって v のクラスを推定する．この決定規則は次のように定式化できる．

$$\begin{aligned}&\text{if } \mathrm{Score}(c,v) = \textstyle\sum_i \mathrm{sim}(v,n_{c,i}) > t_c\\ &\qquad \text{then assign } c \text{ to } v\end{aligned} \quad (4)$$

ここで $n_{c,i}$ はクラス c に属する k 近傍事例の一つ，t_c はクラスごとの閾値，$\mathrm{sim}(v,n_{c,i})$ は引数間の cos 類似度を返す関数である．閾値 t_c は，任意の評価指標を最大化するように訓練データを用いて設定できる．

本研究では，次のように k 近傍事例のうちクラス c に属する数（$|n_c|$ で表わす）で cos 類似度を乗じるように評価関数に変更を加えた．

$$\begin{aligned}&\text{if } \mathrm{Score}(c,v) = \textstyle\sum_i \mathrm{sim}(v,n_{c,i}) \times |n_c| > t_c\\ &\qquad \text{then assign } c \text{ to } v\end{aligned} \quad (5)$$

これによって，k 近傍中に多く含まれるクラスは高スコアを得やすくなる．予備実験では，この修正によってわずかではあるものの 2%程度の精度向上が見られた．

3. 評　　価

3.1 評価用データ

提案する枠組みの有効性を評価するため，TREC Genomics トラックのデータ [1] を利用して評価実験を行った．このデータは訓練用として 374 件，テスト用として 378 件の全文論文（full-text article）から成る．それぞれの論文は一つ以上の遺伝子と組になっており，それぞれの遺伝子は一つ以上の GO ドメインあるいはラベル NEG が MGI の専門家によって与えられている．三つ組み ⟨article, gene, class⟩ の総数は，訓練データとテストデータでそれぞれ 1,661 件（正事例が 589，負事例が 1,072）と 1,077 件（正事例が 495，負事例が 582）である．

表 1　TREC の公式結果と提案手法による結果の比較
Table 1　The TREC official results and our results for GO domain code annotation.

		Prec	Recall	F_1
TREC	Best	0.441	0.769	0.561
	Worst	0.169	0.133	0.149
	Mean	0.360	0.581	0.382
提案手法	TFIDF	0.549	0.642	0.592
	TFCHI_1	0.480	0.630	0.545
	TFCHI_2	0.508	0.731	0.600

評価実験に先立ち，次のようにデータの前処理を行った．まず，遺伝子名はしばしばギリシャ文字を含むので，それらを表わす文字エンティティ（例えば `&agr;`）は，前もって対応する英語の綴りに変換した（例えば `alpha`）．また，セクションタイトルも GO アノテーションに有用である可能性があるので，セクションを構成する全ての段落にそのタイトルを付加した．さらに，図表のキャプションは論文中でそれらの図表を初めて参照する箇所に挿入した．

3.2 評 価 指 標

TREC のデータを用いた他の実験報告との比較を容易にするため，同一の評価指標である F_1 値を用いた．F_1 値は次のように，再現率（R）と適合率（P）の調和平均として定義される．

$$\begin{aligned}P &= \frac{\text{システムが付与した正しいクラス数}}{\text{システムが付与したクラス総数}}\\ R &= \frac{\text{システムが付与した正しいクラス数}}{\text{MGI によって付与されたクラス総数}}\\ F_1 &= \frac{2\times P\times R}{P+R}\end{aligned} \quad (6)$$

ここでクラスは BP，CC，MF のいずれかである．

4. 結果と考察

4.1 TREC 公式結果との比較

まず，訓練データを用いて k の値と各クラスごとの閾値 t_c を単語重み付け法ごとに F_1 を最大にするように設定した．表 1 に，テストデータにおける TREC の公式結果と提案手法による結果をまとめる．

提案手法の単純さにも関わらず，特に TFIDF と TFCHI_2 において良好な結果が得られた．なお，TREC 公式結果の Best は，TFCHI_2 を用いた本手法によって得られたものである [8]．F_1 値が 0.561 から 0.600 に向上しているのは，素性選択・分類におけるコードの修正による．以下の節では，本論文で提案した GO ドメイン推定の枠組み，特に文書表現に関する構成要

素について，その効果を実験を通して精査していく．

4.2 追加実験

4.2.1 提案枠組み各部の有効性検証

本論文で提案した GO ドメインコード推定の枠組みには，例えばテキスト断片の抽出単位など，種々の恣意的な要素が含まれている．これらの妥当性および効果を検証するため，以下の点に着目して実験を行った．

- 遺伝子名同定：表記の様々な違いに対応するため，提案手法では曖昧単語列一致によって遺伝子名を同定した．その効果を検証するため，完全単語列一致によって遺伝子名同定を試みた．
- 遺伝子名辞書：自動構築した遺伝子名辞書による遺伝子名拡張の効果を検証するため，辞書を用いずに実験を行った．
- Glossary と Keyword フィールド：入力論文の当該フィールドを用いた遺伝子名拡張が効果的であったかを検証するため，これらの情報を利用せずに実験を行った．
- MeSH 索引語：GO ドメインの推定における MeSH の有用性を検証するため，MeSH 索引語を素性として利用せずに実験を行った．
- テキスト抽出単位：段落がテキスト断片の抽出単位として妥当であったかを，他の抽出単位（下記）と比較して検証した．
  - 遺伝子名を含む文のみ（以下 G で表わす）
  - G に加え，それに続く文（G+S）
  - G+S に加え，それに先行する文（P+G+S）
  - 遺伝子名出現の有無に関わらず論文全体（ART）

G+S と P+G+S に関しては，抽出の際に段落の境界を越えないようにした．なお，段落を抽出単位とする本研究の方法は，P+G+S と ART の中間にあると考えられる．

4.2.2 実験結果

表 2 に訓練データとテストデータにおける結果を示す．kNN を適用する際，訓練データに関しては処理対象以外の論文をラベル付き事例として用い，テストデータに関しては訓練データをラベル付き事例として用いた．最下段の「デフォルト」は表 1 の TFCHI_2 に対応する．F_1 値が異なるのは，閾値に依存しない比較を行うために，この実験では t_c を評価データ上で決定したことによる．また，平均値に続く比率はデフォルトと比較したときの性能変化を示す．

表 2　異なる設定による実験結果（F_1 値）
Table 2　Results for alternative settings in F_1.

設定		訓練	テスト	平均
遺伝子名同定	完全一致	0.354	0.411	0.383 (−34.8%)
遺伝子名辞書	不使用	0.416	0.470	0.443 (−24.5%)
Glossary & Keyword	不使用	0.543	0.635	0.589 (+0.3%)
MeSH	不使用	0.543	0.625	0.584 (−0.5%)
抽出単位	G	0.525	0.613	0.569 (−3.1%)
	G+S	0.532	0.619	0.576 (−2.0%)
	P+G+S	0.535	0.620	0.578 (−1.6%)
	ART	0.491	0.620	0.556 (−5.3%)
デフォルト		0.543	0.631	0.587

a）遺伝子名同定

完全単語列一致を用いた場合，訓練データ・テストデータのいずれにおいても分類精度が大きく低下した．この結果は，学術論文中ではデータベースで用いられる標準的な遺伝子名とは若干異なる名称が広く使われていることを示している．よって，本論文で用いたような柔軟な遺伝子名同定の方法が重要となる．ただし，この方法の問題点として，本来は遺伝子名ではない単語列を遺伝子名と誤認識してしまうことが予想される．しかし，その影響は式 (1) の Overlap スコアに関する閾値を適切に設定することで抑制できる．本研究における実験では，閾値 0.3 が常に最適であった．

b）遺伝子名辞書

辞書による遺伝子名拡張を行わなかった場合，上記と同様に，訓練データ・テストデータのいずれにおいても分類精度が大きく低下した．これは遺伝子名辞書の有効性を実証するものであり，論文に関連付けられている遺伝子が既知の場合は，本研究のように自動構築した辞書でも十分な効果が得られることを示している．

c）Glossary と Keyword

予想に反し，遺伝子名拡張における`<glossary>`と`<keyword>`フィールドの利用は GO ドメイン推定に有効ではなかった．さらに詳細に実験データを分析した結果，これらのフィールドが遺伝子名の同義語を含むケースはごく少数に限られており，結果として分類精度にほとんど影響を及ぼさなかったことが明らかになった．具体的には，実験データを構成する 882 件の論文のうち，遺伝子名とその同義語が定義されているケースはわずか 15 件（1.7%）であった．

d）MeSH 索引語

MeSH 索引語を利用した場合（デフォルト）も利用しなかった場合（MeSH 不使用）も F_1 値はほとんど

変化しなかった．しかし，さらに両者の結果を比較したところ，適合率はデフォルトで 0.503，MeSH 不使用で 0.509 とほとんど変わらないのに対し，再現率はデフォルトで 0.847，MeSH 不使用で 0.808 と前者が約 5%高い値を示していた．つまり，MeSH 索引語を素性として取り込んだことで，わずかに誤推定が増加したものの，5%増の事例に関して正しいクラスを推定できたことが分かる．適合率の低下は，MeSH が遺伝子ではなく論文単位に付与されていることに起因すると推測される．

e） テキスト抽出単位

この実験では，4 つの異なる抽出単位を比較した（表 2 で G から ART に進むに従ってより大きな抽出単位となっている）．結果を見ると，G から P+G+S に進むにつれ F_1 値が向上し，ART で（論文全体を利用したとき）減少する傾向が観察できる．なお，デフォルトはいずれの場合よりも高い精度を示しており，抽出単位として妥当であったことが分かる．しかしながら，テストデータ単独で見ると ART も高い F_1 値を得ており，遺伝子名の出現箇所に注目した枠組みが必ずしも効果的ではない可能性を示唆している．これには次の理由が考えられる．(a) 入力遺伝子名が論文中に遍在するため，段落だけを抽出しても論文全体を利用してもあまり変化がない，(b) 入力論文で言及されている遺伝子数が少なく，遺伝子ごとの文書表現をした効果が薄い，あるいは (c) 遺伝子数が複数でも同じ GO ドメインが付与されることが多い．訓練データとテストデータを比較することでこれらの可能性を検討したものの，両者に目立った違いは見られなかった．

続いて，段落の利用が効果的であったか，さらに詳細な比較を行うため，デフォルトと ART について再現率と適合率の関係を調べた．結果を図 3 に示す．上の二つのグラフがテストデータにおける結果，下が訓練データにおける結果を示す．訓練データよりは効果が小さいものの，テストデータにおいても段落を抽出した場合（デフォルト）の方が広い範囲で高い適合率（precision）を得ていることが観察できる．

4.2.3 セクション単位の重要性評価

本論文で提案した枠組みは，段落を抽出単位に用いているという点では文章の論理構造を利用していると言える．しかし，例えばセクションなどの論文に特徴的な論理構造までは考慮していない．異なるセクションは論文の異なる側面を論じているため，このような情報は GO アノテーションに有用である可能性がある．

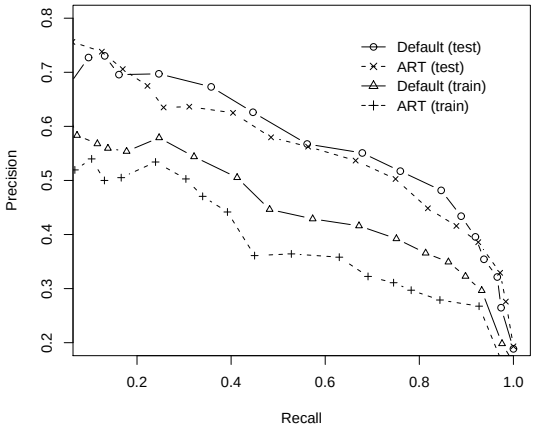

図 3 デフォルトと ART における再現率と適合率の関係
Fig. 3 Relation between recall and precision for Default and ART.

そこで，GO ドメイン推定における個々のセクションの貢献を実験的に調査した．なお，ここでは Abstract，Introduction，Procedures，Methods，Results セクションに注目した．Conclusion は Results と共に記述されることが多いので，Results として扱った．

個々のセクションだけを段落抽出に用いたときの GO ドメイン推定の精度を図 4 に示す．論文タイトルだけを用いた場合，MeSH 索引語だけを用いた場合，全てのセクションを用いた場合（All，表 2 のデフォルトに相当）も参考に示す．棒上の数字は All に対する比率を表している．この実験によると，Results セクションだけを用いた場合でも論文全体を用いた場合と非常に近い結果（97%）が得られることが分かった．一方，Methods セクションは，タイトルのみ利用した場合と比較しても有用な情報を含んでいなかった．なお，テストデータ上でも同様の傾向が見られた．

4.3 Triage タスク

4.3.1 概　　要

本論文で提案した GO ドメイン推定の枠組みは，Genomics トラックの Annotation タスクを念頭に置いている．しかし，この枠組みはもう一方の Triage タスクにも同様に適用可能である．Triage タスクは，入力論文が GO アノテーションの対象となるような実験的記述を含んでいるか否か，すなわち positive か negative を判定する問題であり，一般のテキスト分類問題と同様に考えられる．

テキスト分類の観点から見ると，前節で議論した

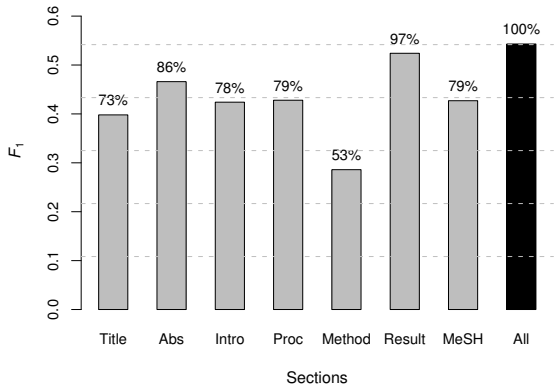

図 4　個々のセクションを単独で用いたときの結果（訓練データを利用）
Fig. 4　Results produced by individual sections on the training data.

GO ドメイン推定と Triage タスクとの違いは，前者が論文と遺伝子の組を入力とするのに対し，後者は論文だけを入力とする点にある．入力が遺伝子を含まないため，Triage タスクでは特定の遺伝子に関する記述だけを同定・抽出する必要はない．しかしながら，Triage タスクも遺伝子に関する判定であるため，本論文で提案した枠組みが有効であると考えられる．そこで，何らかの遺伝子名が出現する段落だけを遺伝子名認識システム YAGI [9] により同定し，それらの段落を抽出して文書表現に用いる．

4.3.2　実験データと評価指標

実験データには Triage タスクのデータを用いる．このデータは訓練用の 5,837 件，テスト用の 6,043 件の全文論文から成り，それぞれ 375 件の正事例と 5,462 件の負事例，420 件の正事例と 5,623 件の負事例を含む．GO ドメイン推定の場合と同様に，訓練データをパラメタの最適化およびテストデータ分類の際のラベル付き事例として用いた．

評価指標には次式で定義される U_{norm} を用いた．

$$U_{norm} = \frac{20 \times TP - FP}{20 \times (TP + FP)} \qquad (7)$$

ここで TP と FP はそれぞれ正しく POS と判断された事例数，誤って POS と判断された事例数を表わす．

4.3.3　結果と考察

表 3 に結果を示す．単語の重み付けに TFCHI_1 を用いた場合，TREC の Best と同一の精度を達成することができた [10]．一方，TFIDF は TFCHI_1 と比較して 30%以上低い U_{norm} を示した．この一因は，MeSH 索引語 `Mice` にある．Dayanik ら [10] は，`Mice` が付与された論文だけを POS に分類するという単純な規則によって Best と同等の結果（$U_{norm} = 0.640$）が得られると報告している．しかし，`Mice` は実験データにおいて高頻度語であるために IDF が低下し，TFIDF では適切な重みを与えられなかったと考えられる．

表 3　Triage タスクにおける TREC の公式結果と提案手法による結果の比較
Table 3　The TREC official results and our results for the triage task.

		Prec	Recall	U_{norm}
TREC	Best	0.157	0.888	0.651
	Worst	0.200	0.014	0.011
	Mean	0.138	0.519	0.330
提案手法	TFIDF	0.112	0.752	0.455
	TFCHI_1	0.160	0.883	0.651
	TFCHI_2	0.137	0.826	0.567

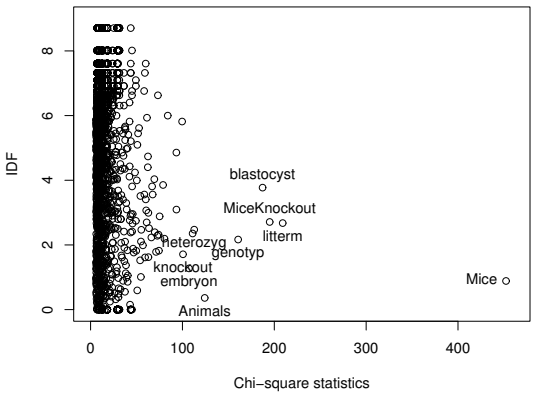

図 5　χ^2 と IDF の散布図
Fig. 5　A scatter plot for χ^2 and IDF.

一方，TFCHI はクラス間の語の分布を考慮することで特定のクラスに集中的に現れるような語に対しても高い重みを付与することができる．IDF と χ^2 の値を比較するため，図 5 に両者の散布図を示す．なお，図中で先頭文字が大文字になっているのは MeSH 索引語である．高い χ^2 を持つ語，例えば `Mice`，`Animals`，`embryon`（embryonic の語幹）は必ずしも高い IDF を持っていないことが観察できる．これは，IDF と χ^2 は高い正の相関を示すというテキスト分類の経験的知見 [5] と矛盾する．興味深いことに，Triage タスクのデータでは両統計量の相関係数は -0.13 であった．この結果は，テキスト分類で一般的に利用されている TFIDF は実験データの特性によって必ずしも適切ではないことを示している．

一方で，TFCHI_1 の高い精度は，図 5 に見られる `Mice` の突出して高い χ^2 のみによって達成された可能性もある．そこで，テストデータから `Mice` を完全に除去した架空のデータを作成し，この上で同じ実験を行った．結果は U_{norm} で 0.548 と表 3 の TFIDF を大きく上回り，TREC 参加者と比較しても 2 位チームと同等の結果であった．

5. 関連研究

GO ドメイン推定（Annotation タスク）と Triage タスクに関して，代表的な研究についてその概要を示し，本研究との関連を述べる．

GO ドメイン推定に関しては，Settle ら [11] がベイズ（NB）分類器と最大エントロピー（ME）モデルを多段的に用いた枠組みを提案している．彼らは，論文の論理構造（セクション）に注目し，あらかじめ定義した 6 つのセクション型のそれぞれについて NB 分類器を構築し，その結果を ME モデルを用いて統合した．ME モデルは，それぞれのセクション型とクラスに応じた重み付けを行う．NB 分類器で用いられた素性は，論文の本文から抽出した単語，統語パターン，そして重要語である．統語パターンとは，主語と直接目的語の組であり，訓練データから自動的に収集される．重要語は高い χ^2 を持つ単語 n グラム（$1 \leqq n \leqq 3$）である．さらに，訓練事例数を増やすため，BioCreAtIvE(注1) などのデータを流用した．以上述べた枠組みによって Settle らは F_1 値で 0.514 の結果を得た．これは本研究で得た値よりも 14%低い．この違いは，Settle らの手法では遺伝子名拡張や曖昧単語列一致等を用いていないためだと考えられる．

Triage タスクに関しては，Dayanik ら [10] がベイズ・ロジスティック回帰分析（BLR）モデルを適用し，TREC 参加者中，最も高い分類精度を得た．この枠組みでは，文書表現に入力論文の要約とそれに付与された MeSH 索引語を用い，単語重みに TFIDF を利用する．これを基本に，まず MeSH 索引語 `Mice` が付与されていない論文は NEG，付与されている論文は BLR によって分類するという二段の枠組みによって，$U_{norm} = 0.641$ を得た．本論文では，TFIDF による重み付けは Triage タスクにおいて悪影響を及ぼすことを示したが，Dayanik らの枠組みでは第一段階のフィルターの効用により最終的に良い結果が得られている．

(注1)：`http://biocreative.sourceforge.net/`

6. あとがき

本論文では，テキスト分類の手法を GO ドメイン推定に適用し，その枠組みと評価実験について詳説した．提案手法では，まず，それぞれの入力 $\langle \mathrm{article}, \mathrm{gene} \rangle$ について当該遺伝子を含む記述を抽出，これに基づいて GO ドメインの推定を行った．遺伝子名の同定には，既存のデータベースから自動的に構築した遺伝子名辞書を用い，微妙な表記の揺れを吸収するために辞書エントリと入力文を正規化し，さらに曖昧単語列一致によって不規則な表記の違いに対応した．こうして抽出されたテキスト断片を基に素性選択を行い，そこで算出したカイ二乗値を単語の重み付けに再利用した．実データを用いた評価実験の結果，提案手法によって従来研究と同等以上の性能が得られた．提案手法を詳細に分析したところ，遺伝子名拡張と曖昧単語列一致を併用して遺伝子名同定に利用することが GO ドメイン推定に最も重要であることが分かった．また，セクションによって GO ドメイン推定に関する重要度が異なり，Results セクションが最も有用な情報を含むことが明らかになった．さらに，提案手法を Triage タスクに適用し，特に教師付き重みを用いた際に高い分類精度が得られた．

今後の課題としては，セクション情報や遺伝子名の周辺文脈の効果的な利用が挙げられる．このような情報は，素性として，あるいは素性の重みとして現在のモデルに取り込むことができる．また，より発展的な課題として，遺伝子名を与えずに論文のみを入力として GO アノテーションを行うことを考えている．

文　献

[1] W. Hersh, R. T. Bhuptiraju, L. Ross, L. Ross, A. M. Cohen and D. F. Kraemer: “TREC 2004 genomics track overview”, Proceedings of the 13th Text REtrieval Conference (TREC) (2004).

[2] A. S. Schwartz and M. A. Hearst: “A simple algorithm for identifying abbreviation definitions in biomedical text”, Proceedings of the Pacific Symposium on Biocomputing (PSB), Vol. 8, pp. 451–462 (2003).

[3] S. Egorov, A. Yuryev and N. Daraselia: “A simple and practical dictionary-based approach for identification of proteins in MEDLINE abstracts”, Journal of the American Medical Informatics Association, **11**, 3, pp. 174–178 (2004).

[4] J. B. Lovins: “Development of a stemming algorithm”, Mechanical Translation and Computational Linguistics, **11**, pp. 22–31 (1968).

[5] Y. Yang and J. O. Pedersen: “A comparative study on feature selection in text categorization”, Proceedings of the 14th International Conference on Machine Learning, pp. 412–420 (1997).
[6] F. Debole and F. Sebastiani: “Supervised term weighting for automated text categorization”, Proceedings of SAC-03, 18th ACM Symposium on Applied Computing, pp. 784–788 (2003).
[7] Y. Yang and X. Liu: “A re-examination of text categorization methods”, Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, pp. 42–49 (1999).
[8] K. Seki, J. C. Costello, V. R. Singan, and J. Mostafa: “TREC 2004 genomics track experiments at IUB”, Proceedings of the 13th Text REtrieval Conference (TREC) (2004).
[9] B. Settles: “Biomedical named entity recognition using conditional random fields and rich feature sets”, Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA) (2004).
[10] A. Dayanik, D. Fradkin, A. Genkin, P. Kantor, D. D. Lewis, D. Madigan and V. Menkov: “DIMACS at the TREC 2004 genomics track”, Proceedings of the 13th Text REtrieval Conference (TREC) (2004).
[11] B. Settles and M. Craven: “Exploiting zone information, syntactic rules, and informative terms in gene ontology annotation of biomedical documents”, Proceedings of the 13th Text REtrieval Conference (TREC) (2004).

（平成 xx 年 xx 月 xx 日受付）

関　和広

2000 図書館情報大学図書館情報学部卒．2002 同大大学院情報メディア研究科博士前期課程修了．2006 米国インディアナ大学大学院図書館情報学研究科博士課程修了．2006 神戸大学大学院自然科学研究科助手．2007 同大自然科学系先端融合研究環助教．知的なテキスト情報システムの研究に従事．

モスタファ ジャビド

1994 米国テキサス大学大学院博士課程修了．1994 米国インディアナ大学大学院図書館情報学研究科助教授．2000 同大准教授．

Abstract Gene Ontology (GO) was developed for describing gene functions to provide a common vocabulary to search across different model organism databases. As a first step toward automatic GO annotation, this study aims to assign GO domains given an article discussing some aspects of genes. We approach the task with careful consideration of the specialized terminology and pay special attention to various forms of gene synonyms. We extract the words around the spotted gene occurrences and feed them to a variant of kNN with supervised term weighting schemes for GO domain annotation. Experimental evaluation demonstrates that our approach compares favorably with the top-performing groups in the TREC official evaluation.

Key words Ontology, Classification, Gene Function, Gene Name Identification, Supervised Term Weighting