共起・連接頻度グラフに基づいた日本語略語展開語候補の生成

篠原（山田）恵美子[†]　荒牧英治[††§]　大江和彦[†]
三浦康秀[‡]　外池昌嗣[‡]　大熊智子[‡]　増市博[‡]
[†]東京大学医学部附属病院　[††]東京大学知の構造化センター　[§]科学技術振興機構さきがけ
[‡] 富士ゼロックス株式会社

emiko-tky@umin.net, eiji.aramaki@gmail.com, kohe@hcc.u-tokyo.ac.jp,
{yasuhide.miura, masatsugu.tonoike, ohkuma.tomoko, hiroshi.masuichi}@fujixerox.co.jp

1　はじめに

我々は電子カルテ等の日本語医療テキストを対象とした言語処理研究を行っている．医療テキストでは略語が多く使われており，曖昧性の解消・名寄せのためにはこれを元の形に戻すこと（略語展開）が必要である．略語展開は展開語リストの中から出現文脈の類似するものを選択するという語義曖昧性解消の枠組みで行われることが多い [7]．展開語リストの生成も多く試みられている [2][4][5]．しかし我々の対象である医療分野では略語・展開語ともに生産的であるため，網羅的な展開語リストの仮定が難しい．我々が調査したところ，医療分野の略語集[1]に収載されている日本語略語で，展開語が医学辞書[2][3]に載っているものは3割程度にとどまった．また展開語リストの拡充や出現文脈の獲得に必要となる十分なサイズのコーパスもない．従って，展開語リストを仮定しない略語展開手法が望まれる．

展開語リストを仮定しない手法としては，ルールに基づいた略語展開 [8] が挙げられるが，「血球計算-血算」「モルヒネ-モヒ」のような変則的な略語生成に対応できないという欠点がある．他に，展開語となる語から略語を推定する方法 [3] が提案されているが，展開語が生産的な場合に扱いが困難である．以上のような背景から，本研究では略語を構成する文字を被覆する単語の組み合わせ全てを展開語候補とした略語展開に取り組んでいる．

図1の例では，略語「心カテ」に対し展開語候補は $2 \times 2 \times 2 = 8$ 通りとなる．実際には，一文字あたりの候補語数は数百のオーダとなるため，展開語の候補数は少なくとも数千，多ければ数百万のオーダとなる．

[1]カルテ＆レセプト略語 16000. 医学通信社
[2]南山堂 医学大辞典
[3]医学書院 医学大辞典

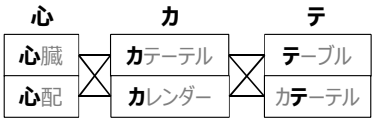

図 1: 略語展開

即ち，展開語リストが無いという前提で略語展開をするためには，この候補語の中から (1) 意味を成す複合語を出力候補として残し，その中から (2) 文脈に沿ったものを選択する技術が必要となる．本稿では (1) に着目する．以下,「意味を成す複合語」の意で「**正例語**」という表現を使う．

正例語か否かを判別する方法として，コーパス中に出現する名詞連続の左右文脈から複合語らしさを表すスコア [9]，複合語およびその構成語の訳語のコーパス出現頻度に基づいたスコア [6] 等が提案されている．また素朴にコーパス中の出現頻度をスコアとして利用する方法も考えられる．しかしいずれの手法でも大量の候補についてコーパス中の出現回数やそれぞれにおける周辺文脈を調べる必要があり，(a) 時間がかかる (b) コーパスが必要であるという問題がある．また，入力が与えられる前に予めスコアを算出しておくのは，候補語が無限に生産可能であることから非現実的である．

一方で，単語対についての情報を持つのは，単語集合を定めてしまえば保持する情報量は一定であるため，現実的な設定といえる．本稿では，図2に示すような単語対の共起・連接頻度グラフに基づいて，ある単語集合が正例語となるか否かを判定する手法を提案する．本手法は原語リストを仮定しない略語展開のみならず，候補語を生成するタスク一般において有用である．また，今回は日本語での評価実験を行ったが，提案手法は言語非依存であり，他言語への適用も可能である．

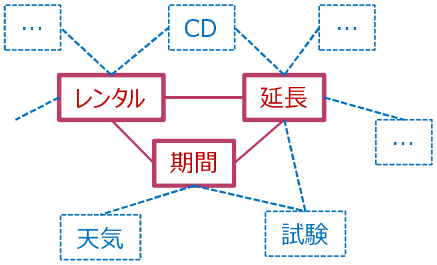

図 2: 単語の連接・共起頻度グラフ．リンクは両端ノードの連接または共起頻度情報を持つ．

2　正例語らしさの表現

2.1　正例語判定に必要な情報

与えられた単語列が正例語か否かの判別方法としては，コーパス中の出現頻度や n-gram 言語モデルが考えられる．しかし本稿では前章で述べた通り単語対の情報のみを用いるため，trigram 以上の出現頻度は使えず，また n-gram 言語モデルでは $n = 2$ となるため長さ 3 以上の列に対しては非常に低い精度となることが予想される．以下に bigram 言語モデルから生成した例を挙げる．

(1) 個人｜情報｜保護（7000 万件）

(2) レンタル｜期間｜延長（4 万件）

(3) オイル｜交換｜日記（2 万件）

(4) 激安｜専門｜分野（0 件）

参考値として括弧内にウェブ上の出現頻度の概数を記載した．(1) は正例語，(4) は正例語でないと言ってよいだろう．(2) は，「一般に何かをレンタルする時には期間が定められており，その期間が延長されることがある」という常識から，意味の成立する文字列であると言えるが，正例語か否かの判断には議論が必要であろう．(3) は意味が成立するか否か不明である．「オイル交換日記」なるものが世界のどこにも存在しない保障はない．あるコミュニティではオイル交換について書かれた日記の存在が常識であるかもしれない．このような存在の有無に関する知識を網羅的に収集するのは難しいが，「構成単語が同一トピックで出現するか否か」がそのような知識の代替になるのではないかと考えた．

上記のことから，正例語判定には 2 つの情報が必要であると考えた．

直接共起性

構成単語が連接もしくは共起していれば複合語となりやすい

トピック性

構成単語が同じトピックに属していれば複合語となりやすい

トピックの表現方法として，本稿では入力単語集合に含まれない語との共起・連接頻度分布を利用することとした．

2.2　正例語らしさの素性化

前節で挙げた 2 つの情報を素性化する方法について述べる．

N 語の単語から成る単語集合を D，共起頻度行列を C，連接頻度行列を S とする．

$$\begin{aligned} D &:= \{w_i\,(1 \le i \le N)\} \\ C &:= \{c_{ij}|w_i\text{と }w_j\text{の共起頻度 }\} \\ S &:= \{s_{ij}|w_i\text{と }w_j\text{の連接頻度 }\} \end{aligned}$$

D の部分集合である入力単語集合を W とする．W の元の添え字集合を I とする．

$$W \subset D, \qquad W := \{w_i|i \in I\}$$

W に対し，内部連接頻度行列 S_{in}，内部共起頻度行列 C_{in} および外部連接頻度行列 S_{out}，外部共起頻度行列 C_{out} を以下のように定義する．

$$\begin{aligned} S_{in} &:= \{s_{ij}\,(i \in I, j \in I)\} \\ C_{in} &:= \{c_{ij}\,(i \in I, j \in I)\} \\ S_{out} &:= \{s_{ij}\,(i \in I, j \notin I)\} \\ C_{out} &:= \{c_{ij}\,(i \in I, j \notin I)\} \end{aligned}$$

前節の直接共起性は入力単語対の内部連接頻度分布および内部共起頻度分布の特徴に，トピック性は入力単語間での外部連接頻度分布および外部共起頻度分布の類似度に対応する．次節の実験では前者として S_{in} と C_{in} の対角成分を除く上三角成分の最大値・最小値・平均値を，後者として S_{out} と C_{out} の行ベクトル間のコサイン類似度を用いた．図 3 に素性の例を示す．

3　実験

3.1　実験データの準備

実験を行うためには学習・テストデータとして正例・負例を用意する必要があるが，負例となるデータは不

$$
\begin{aligned}
\text{単語集合 } D &= \{\text{天気}, \ldots, \text{レンタル}, \text{試験}, \ldots, \text{期間}, \ldots, \text{延長}, \ldots, CD\} \\
\text{入力単語集合 } W &= \{\text{レンタル}, \text{期間}, \text{延長}\} \\
\text{直接共起性} &: f(s_{\text{レンタル}}, s_{\text{期間}}, s_{\text{延長}}),\ f(s_{\text{レンタル, 期間}}, s_{\text{期間, 延長}}, s_{\text{延長, レンタル}}) \\
\text{トピック性} &: f(cos(\vec{s}_{out,\ \text{レンタル}}, \vec{s}_{out,\ \text{期間}}), cos(\vec{s}_{out,\ \text{期間}}, \vec{s}_{out,\ \text{延長}}), cos(\vec{s}_{out,\ \text{延長}}, \vec{s}_{out,\ \text{レンタル}})) \\
f &:= \{f | max, min, average\}
\end{aligned}
$$

図 3: ＜レンタル, 期間, 延長＞の素性化

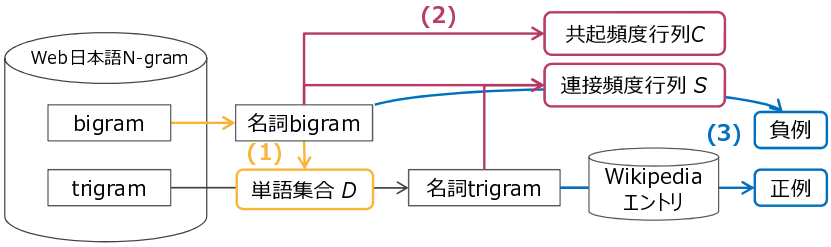

図 4: 実験データの作成手順．(1) 名詞のみで成る bigram から単語集合を構築，(2) 名詞 bigram から頻度行列を構築，(3) 名詞 trigram かつ Wikipedia エントリを正例，bigram 言語モデル（共起頻度行列に基づく）から生成した trigram を負例とする．

適格な語集合であり，一般に存在するものではない．本稿では [1] と同様に，bigram 言語モデルから生成した単語列を負例とすることとした．

実験データの作成は，(1) 単語集合 D の構築，(2) 共起／連接頻度行列 C ／ S の構築，(3) 正例・負例の作成の 3 段階から成る（図 4）．利用したリソースは Web 日本語 N グラムと Wikipedia [4]見出し語の二点である．

(1) 単語集合 D の構築

bigram を形態素解析[5][6]し，名詞のみから成る bigram を抽出して名詞 bigram とした．得られた名詞 bigram に出現する語を収集し，単語集合 D とした．また trigram のうち，D に含まれる語のみから成るものを抽出し，名詞 trigram とした．

(2) 共起頻度行列 C，連接頻度行列 S の構築

名詞 n-gram($2 \leq n \leq 3$) での共起頻度から C を，名詞 bigram での共起頻度から S を構築した．

(3) 正例・負例の作成

正例は名詞 trigram のうち Wikipedia の見出し語であるもの，負例は名詞 bigram を言語モデルとして作成した trigram とし，それぞれ 1000 語ずつを実験に用いた．正例および負例の一部を以下に示す．

正例　個人-情報-保護，開発-途上-国，消費-者-金融

負例　中華-代表-ドーハ，影武者-日本-百科，茨城-風-生け花，人形-焼-ゼイタク，針刺し-防止-靴

3.2 評価

正例・負例を用い，TinySVM[7]による「正例語である／ない」の分類精度を 2 分割交差検定により計測した．$S_{in}, C_{in}, S_{out}, C_{out}$ が判別精度に与える影響，およびそれぞれのスコア平均と標準偏差を調べた．ここでスコアとは SVM の分離平面からの距離である．

結果を表 1 に示す．S_{in}, C_{in} のいずれかを用いた場合に高い性能を示し，S_{out}, C_{out} の効果は分からなかった．S_{in} のみを用いた場合と S_{in}, S_{out} を用いた時には正例のスコア，C_{in} のみを用いた場合と C_{in}, C_{out} を用いた時には負例のスコアについて，それぞれ絶対値が大きい結果となった．また，S_{in}, C_{in} の両方を用いた時，負例のスコアの標準偏差が小さくなっており負例の性質をよく表している素性であることが伺われる．

4 考察

本稿で述べた実験では (1) 判別対象，(2) 入力単語集合のサイズに限界がある．

(1) 本来判別したいのは「正例語か否か」であるが，実験では入力単語集合が Wikipedia エントリと bigram 言語モデルのいずれであるかを判別している．正例と

[4]Wikipedia は Wikimedia Foundation, Inc. の登録商標.
[5]MeCab. http://mecab.sourceforge.net/
[6]UniDic. http://www.tokuteicorpus.jp/dist/
[7]TinySVM. http://www.chasen.org/ taku/software/TinySVM/

表 1: 判別精度 (線形分離)

S_{in}	C_{in}	S_{out}	C_{out}	適合率	再現率	F 値	スコア平均（正例）	スコア平均（負例）
+				1.000	1.000	1.000	2.645±0.572	-1.565±0.261
	+			1.000	1.000	1.000	1.795±0.960	-5.377±1.135
+	+			1.000	1.000	1.000	1.757±1.350	-1.567±0.147
		+		0.882	0.872	0.877	1.293±1.085	-1.125±0.885
			+	0.880	0.869	0.874	1.514±1.297	-1.110±0.890
		+	+	0.885	0.876	0.880	1.588±1.313	-1.210±0.971
+		+		1.000	1.000	1.000	2.628±0.569	-1.556±0.261
	+		+	1.000	1.000	1.000	1.775±0.943	-5.201±1.069
+	+	+	+	1.000	1.000	1.000	1.748±0.353	-1.562±0.145

した Wikipedia エントリは正例語と言えるだろうが，負例は bigram 言語モデルから生成した擬似負例であり，正例を含む可能性がある．したがって適合率は高いほうが望ましいが，再現率はそうであるとは限らない．より精細な評価のためには疑似負例を正例語らしさの観点で分類し，システムによるスコア付けの適切さを調査する必要がある．

(2) 入力単語集合 W のサイズを 3 に固定して実験を行ったが，より大きなサイズでの精度を調査する必要がある．

また，展開語の中に構成要素として現れうる語を網羅していなければ，正しい略語展開は不可能である．実用的な略語展開を見据えるならば，単語集合 D を綿密に設計する必要がある．

5 おわりに

本稿では，語集合が一つの複合語として成り立つか否かについて，集合内での語の共起・連接頻度，および集合外の語との共起・連接頻度に基づいたスコアの算出手法を提案した．疑似負例を用いた trigram についての実験の結果，集合内での連接頻度および共起頻度が精度に大きく影響し，集合外の語との関係は精度に影響を与えなかった．今後，より現実に即した設定での精度評価および trigram よりも長い複合語を対象とした精度評価を行う予定である．

参考文献

[1] Daisuke Okanohara and Jun'ichi Tsujii. A discriminative language model with pseudo-negative samples. In *ACL*, pp. 73–80, 2007.

[2] Naoaki Okazaki, Mitsuru Ishizuka, and Jun'ichi Tsujii. A discriminative approach to japanese abbreviation extraction. In *IJCNLP*, pp. 889–894, 2008.

[3] 村山紀文, 奥村学. Web 情報を利用した確率モデルによる略語推定. 情報学基礎研究会報告, 第 4 巻, pp. 93–100, 2008.

[4] 内海慶, 小町守, 町永圭吾, 前澤敏之, 佐藤敏紀, 小林義徳. 検索クエリログとクリックスルーログを用いた略語の展開候補獲得. 情報処理学会研究報告, 第 4 巻, pp. 1–7, 2010.

[5] 酒井浩之, 増山繁. 略語とその原型語との対応関係のコーパスからの自動獲得手法の改良. 自然言語処理, Vol. 12, No. 4, pp. 1–25, 2005.

[6] 外池昌嗣, 宇津呂武仁, 佐藤理史. ウェブから収集した専門分野コーパスと要素合成法を用いた専門用語訳語推定. 自然言語処理, Vol. 14, No. 2, pp. 33–68, 2007.

[7] 寺田昭, 徳永健伸. 文脈情報を使用した略語の自動復元. 自然言語処理研究会報告, 第 69 巻, pp. 39–45, 2001.

[8] 石井直樹, 平石智宣, 延澤志保, 斎藤博昭, 中西正和. 日本語略語の自動復元. 自然言語処理研究会報告, 第 53 巻, pp. 61–68, 2000.

[9] 中川裕志, 森辰則, 湯本紘彰. 出現頻度と連接頻度に基づく専門用語抽出. 自然言語処理, Vol. 10, No. 1, pp. 27–45, 2003.