バイオ医療専門用語の類義語獲得

奈良先端科学技術大学院大学　自然言語処理学講座　鈴木郁美　原一夫　新保仁　松本裕治

研究目的（タスク）

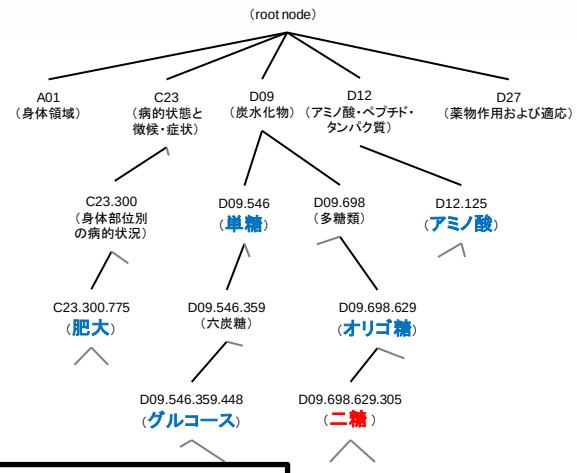
新規専門用語のシソーラス辞書への登録支援を行うシステムの構築.

例）クエリとして 二糖をシステムに与えた場合

(a) システムの出力

類似度ランキング	
1位	オリゴ糖
2位	単糖
3位	アミノ酸
4位	肥大
5位	グルコース

(b) シソーラスへのマッピング



方法

コーパスから文脈情報を取り出し，語の特徴ベクトルを作る

雑誌「蛋白質・核酸・酵素 (PNE)」の例

> 血管内皮細胞は血管内表面を被う単層上皮細胞であり,近年, 組織培養法が確立し, 多方面からの研究が進み, 一躍脚光を浴びるようになった.

文脈情報は，専門用語の隣接語，係り受け，接尾辞とし，
特徴ベクトルの素性として使う.

上皮細胞の素性は，(血管内表面，被う，ある，近年，浴びる，なる，細胞，胞)

隣接語　係り受け　接尾辞

特徴ベクトル：$\mathbf{v}_i (i = 1, \ldots, N)$，$i$：専門用語インデックス

グラフによる語と語の類似度計算

- 専門用語グラフ

  専門用語を節点とし，節点 iと節点 jを結ぶ枝の重み$\mathbf{v}_i \cdot \mathbf{v}_j$ を（内積）として与え，グラフを作成する.

- 隣接行列: A

  ij 要素を$\mathbf{v}_i \cdot \mathbf{v}_j$として持つ行列である.

- コサイン類似度: Cos

  節点iと節点jのコサイン類似度は $\dfrac{\mathbf{v}_i \cdot \mathbf{v}_j}{\|\mathbf{v}_i\| \|\mathbf{v}_j\|}$ により与えられる.

- フォンノイマン拡散カーネル行列: VN　$\mathbf{VN} = \mathbf{A}(\mathbf{I} - \beta\mathbf{A})^{-1} = \mathbf{A} + \beta\mathbf{A}^2 + \beta^2\mathbf{A}^3 + \cdots$

- ラプラシアン拡散カーネル行列: LDF　$\mathbf{LDF} = \exp(-\beta\mathbf{L})$

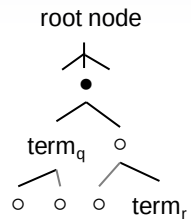
$$= \mathbf{I} + \beta(-\mathbf{L}) + \frac{\beta^2}{2!}(-\mathbf{L})^2 + \frac{\beta^3}{3!}(-\mathbf{L})^3 + \cdots$$

評価方法（正解類似度計算）

クエリの専門用語 $term_q$ に対し，システムが出力する専門用語 $term_r$ とのシソーラス・ツリーにおける正解類似度を，以下に定める.

$term_q$と$term_r$の類似度

$$= 2 \times \frac{term_q \text{ と} term_r \text{の共通祖先節点の深さの最大値}}{term_q \text{ の深さ} + term_r \text{の深さ}}$$



例: 上図における類似度は，2×1÷（2+3）＝0.2

類似度ランキング	専門用語	正解類似度	重み
1位	オリゴ糖	0.86	1.00
2位	単糖	0.33	0.50
3位	アミノ酸	0.00	0.33
4位	肥大	0.00	0.25
5位	グルコース	0.25	0.20

結果

ランク1位出力の評価スコア．正解類似度の平均値の比較.

クエリ		手法		
PNEでの登場回数	クエリ数	Cos	VN	LDF
2回以下	253	0.228	0.231	**0.255**
3回以上	330	**0.324**	0.318	0.257

ランク1~5位出力の評価スコア．正解類似度の平均値の比較.

クエリ		手法		
PNEでの登場回数	クエリ数	Cos	VN	LDF
2回以下	253	0.213	0.214	**0.225**
3回以上	330	**0.265**	0.262	0.214

考察

- 登場回数が少ないクエリは文脈から抽出できる情報が少ない.
- グラフを辿って間接的に文脈を獲得する”拡散させる”VN, LDFがCosより高いスコアを得た.
- VN(隣接行列ベース)よりLDF(ラプラシアン行列ベース)のほうがよいスコアを得た.
- 一般的な専門用語をランキングの上位に出しにくくするというラプラシアン行列の性質を反映したものと示唆される.

謝辞

京都大学の金子周司先生には，ライフサイエンス辞書のシソーラスを提供していただき，深く感謝します.

フォンノイマン拡散カーネル，ラプラシアン拡散カーネルを使う利点

- コサイン類似度（隣接行列）

  共通する文脈が無いとき，類似度は常にゼロ.
  よって，専門用語など，コーパスでの出現頻度が少ない語は，文脈を共有できないために，本来より類似度が低く計算される可能性がある.

- 拡散

  グラフを辿ることで節点間の類似度を測る.
  直接リンクのない節点間の類似度はゼロにならない.

- ラプラシアン行列

  隣接行列との違いは，セルフループの重み.
  一般的な意味をもつ専門用語の節点を通る経路のスコアをディスカウントする.

(c) 専門用語の特徴ベクトル

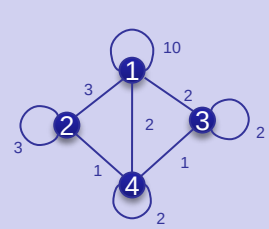
$$\mathbf{v}_1 = \{1,1,1,1,1,1,1,1,1,1\}$$
$$\mathbf{v}_2 = \{1,1,1,0,0,0,0,0,0,0\}$$
$$\mathbf{v}_3 = \{0,0,0,0,0,0,0,0,1,1\}$$
$$\mathbf{v}_4 = \{0,0,1,0,0,0,0,0,1,0\}$$

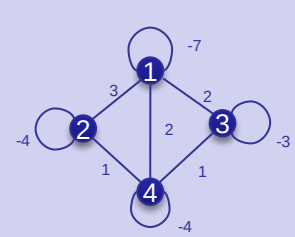
(e) 拡散する

$$\mathbf{A}^2 = \begin{bmatrix} 117 & 41 & 26 & 33 \\ 41 & 19 & 7 & 13 \\ 26 & 7 & 9 & 10 \\ 33 & 13 & 10 & 22 \end{bmatrix}$$

(d) 専門用語グラフ

$$\mathbf{A} = \begin{bmatrix} 10 & 3 & 2 & 2 \\ 3 & 3 & 0 & 1 \\ 2 & 0 & 2 & 1 \\ 2 & 1 & 1 & 4 \end{bmatrix}$$



(f) ラプラシアン行列

$$-\mathbf{L} = \begin{bmatrix} -7 & 3 & 2 & 2 \\ 3 & -4 & 0 & 1 \\ 2 & 0 & -3 & 1 \\ 2 & 1 & 1 & -4 \end{bmatrix}$$