〈コーパスに出現するライフサイエンス分野の専門用語の同定とシソーラスへのマッピング〉

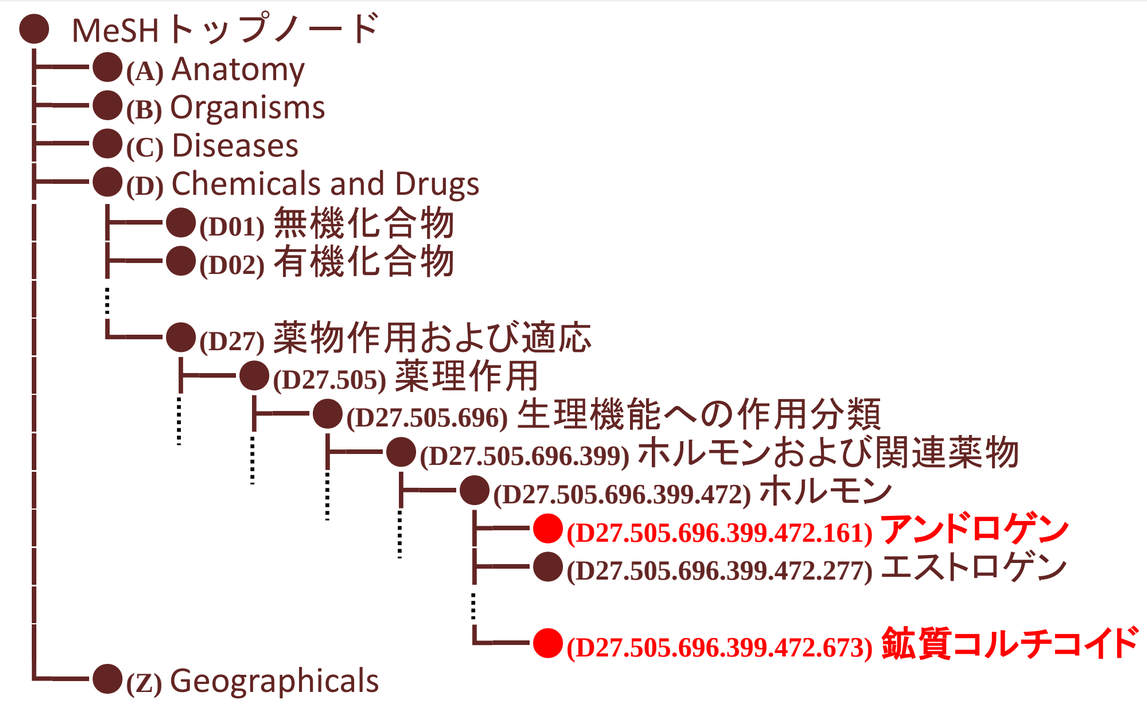
係り受け木を利用した専門用語の類似度計算方法とそのシソーラス拡張への応用

奈良先端大学・情報科学研究科・自然言語処理学講座
原 一夫　鈴木 郁美　新保 仁　松本 裕治

シソーラス

シソーラスとは，単語の上位下位関係をツリーで表した辞書．

例：MeSHシソーラス



例えば，**アンドロゲン(androgen)** と**鉱質コルチコイド(mineralocorticoids)** は，ホルモンを上位語とする兄弟関係にある．

シソーラス拡張

目的：　辞書編集者がシソーラスに新しい単語を追加登録するのを支援する．

手順：　コーパスでの出現文脈を用い，新しい単語と登録済み単語の類似度を計算し，類似度の高い単語の近くに新しい単語を配置することを推薦する．

単語の類似度を計算する標準的な手法（コサイン類似度による手法）

ウインドウサイズWで周辺単語を抽出し，特徴ベクトルを作る．

例文S: Androgen receptor blockade might be potentiated by a reduction of serum **androgens** .

例文S': These results can not be explained by receptor down-regulation due to higher levels of **mineralocorticoids** in PIH .

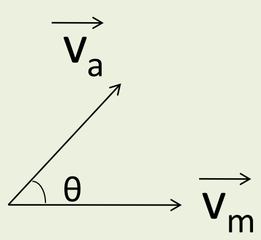
androgens の特徴ベクトル $\vec{V_a}$ =

1	receptor
1	blockade
1	might
1	be
1	potentiated
1	by
1	a
1	reduction
1	of
1	serum
0	explained
0	down-regulation
0	due
0	to
0	higher
0	levels
0	in
0	PIH

mineralocorticoids の特徴ベクトル $\vec{V_m}$ =

1	receptor
0	blockade
0	might
1	be
0	potentiated
1	by
0	a
0	reduction
0	of
0	serum
1	explained
1	down-regulation
1	due
1	to
1	higher
1	levels
1	in
1	PIH

コサイン類似度：　2つの単語の類似度を，2つの単語に対応する特徴ベクトルが成す角度θのコサインと定義する．

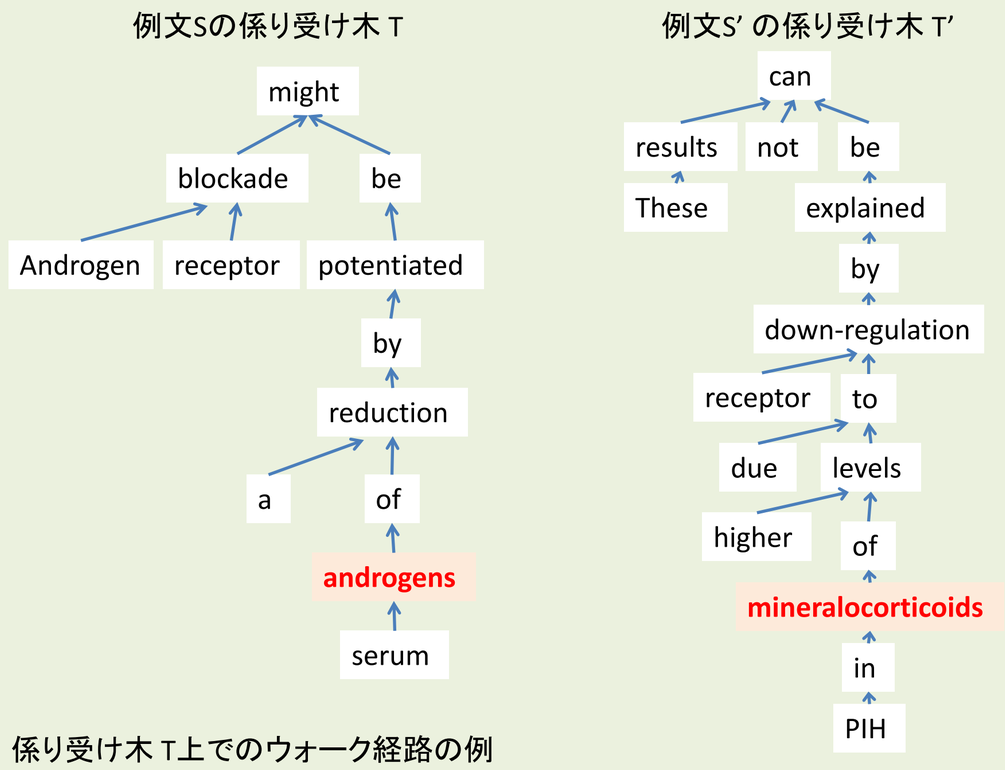


本研究で検証する仮説

文の係り受け木構造の情報を用いれば，より正確な単語の類似度を計算できるのではないか．

単語類似度計算の提案手法

注目する2つの単語をそれぞれ含む2つの係り受け木 T，T'において，注目する単語をそれぞれ出発点とするランダムウォークを並行して行う．



係り受け木 T上でのウォーク経路の例

androgens → of → reduction → a → reduction → by

係り受け木 T'上でのウォーク経路の例　（点線は，係る方向と逆向きのウォーク）

mineralocorticoids → of → levels ⇢ higher → levels → to

ウォーク経路ペアの類似度を，対応する節点（単語）の類似度，および，対応する枝（係る方向）の類似度の積と定義する．

節点（単語）の類似度は，コサイン類似度で測る．

枝の類似度は，方向（順向き，逆向き）が一致するとき1，一致しないとき 0 とする．

2つの単語の類似度を，全てのウォーク経路ペアの類似度の総和と定義する [Kashima et al. 2003].

評価実験

コーパス：GENIA（1999 MEDLINE アブストラクト）
シソーラス：MeSH
実験対象：GENIAコーパスに出現し，かつ，MeSHアドレスを複数持たない234専門用語
実験方法：1つをクエリとし，残りの233をランキングすることを繰り返す．
パラメタ：ウォークを終了する確率をγとする．なお，節点から隣接節点への遷移は等確率とする．
評価方法：ランキング結果における，正解の最上位ランク（のクエリ平均）を評価指標とする．ここで，正解をクエリの親，兄弟，子とする．
実験結果：表に示す．Nはウォークの長さであり，N＝0は，係り受け木構造を用いないコサイン類似度による手法に相当する．提案手法はコサイン類似度による手法を上回る結果を得ることに成功した．

γ	0.001	0.005	0.01	0.05	0.1	0.2	0.3	0.4	0.5
N＝0	42.48	42.48	42.48	42.48	42.48	42.48	42.48	42.48	42.48
N≦1	40.43	40.43	40.43	40.31	39.70	**38.99**	**39.54**	**40.68**	**41.29**
N≦2	38.49	38.5	38.50	**38.21**	**38.85**	40.72	41.36	41.82	42.06
N≦3	39.17	39.12	38.88	38.34	39.92	41.17	41.68	41.91	42.07
N≦4	38.08	37.79	**37.81**	39.95	40.92	41.47	41.75	41.91	42.06
N≦5	37.94	**37.56**	38.18	40.40	41.14	41.51	41.76	41.91	42.06
N≦6	37.67	38.3	39.31	40.92	41.24	41.53	41.76	41.91	42.06
N≦7	**37.21**	38.73	39.94	41.03	41.26	41.53	41.76	41.91	42.06
N≦8	38.26	39.7	40.17	41.16	41.30	41.53	41.76	41.91	42.06
N≦9	38.9	40.1	40.45	41.19	41.30	41.53	41.76	41.91	42.06
N≦10	39.37	40.26	40.66	41.21	41.30	41.53	41.76	41.91	42.06

(cc) BY Licensed under a Creative Commons表示2.1日本 license (c)2010 原 一夫（奈良先端科学技術大学院大学）