疾患間類似度計算における分散表現の活用手法

大村 舞 †　　松本 裕治 †,§　　奥村 貴史 ‡,§

† 奈良先端科学技術大学院大学 情報科学領域
‡ 北見工業大学 工学部　　§ 理化学研究所 革新知能統合研究センター
†{omura.mai.oz5, matsu}@is.naist.jp,　‡tokumura@mail.kitami-it.ac.jp

1 はじめに

人工知能技術の発展を受けて、人工知能技術の医療への応用が盛んとなっている。しかし、診断アルゴリズムの研究は高度化しているが、医療用人工知能の診断結果を効率的かつ効果的に提示するユーザーインターフェースの研究については、発展が遅れている。たとえば、診断支援システムは患者の呈している症状から可能性の高い疾患をリスト化できるが、リストが長大化しやすく、医師の認知的負荷を高めてきた [5]。もし、とある疾患に着目した際、疾患の単純なリストではなく、その疾患に類似した疾患を合わせて提示することができれば、診断結果をより効率的に提示することが可能となる。

このようなユーザーインターフェースを実現するためには、任意の疾患と疾患との類似度を定量化する手法が必要となる。こうした定量化に際しては、人手で構築された疾患オントロジーなどを用いて小規模な疾患間の類似度を計算する試みは存在した [3]。しかし、類似度計算手法について網羅的に検討した研究は知られていない。そこで我々のグループでは、オントロジーよりも低コストに編纂しうる簡易的な疾患知識ベース [6] を用いて、網羅的な疾患間類似度計算の実現に取り組んできた [11]。しかし、疾患知識ベースの構築には、それでも相当のコストを求められる。

近年、自然言語処理分野において、単語等の概念を分散表現にすることにより、品詞タグ付け、文書分類、機械翻訳、評判分析といったさまざまなタスクにおける性能向上が報告されている [10]。医療分野においても、電子カルテなどから構築された単語の分散表現が、医療的な概念の類似関係の計算 [1] や、疾患のカテゴライズ [2] といったタスクに有効であることが示されてきた。もし、疾患を分散表現化することができれば、内積等を用いることで容易に類似度計算することが可能となることが期待される。

そこで本稿では、単語の分散表現により疾患類似度計算を低コスト化し、また、高精度化することが可能かを検討する。word2vec (Skip-gram モデル) により構築された分散表現は、テキストの文脈を予測するようなベクトルを生成する [4]。この類似する文脈に共起しやすい疾患 (の名称) 同士の距離が、疾患類似度として有効たりうるか、実証的に検討する。次節以降、まずモデル生成に用いたデータ及び精度管理に用いた疾患類似度データについて概要を示す。その後、実験結果について示すと共に、結果に考察を加える。最後に、今後の展望を整理し結語を記す。

2 疾患名の分散表現による疾患間類似度手法

2.1 分散表現に用いるデータ

分散表現データの構築に用いるテキストデータとしては、ウェブより入手可能な英文医学文献のテキストデータを用いた。このデータは、文献 [6] にて構築した疾患データベースの疾患名をクエリとして Google で検索した結果より、上位 35 件のページをクロールしたものである。全部で 156,422 件が含まれており、ここから HTML タグを除外し、抽出した文を小文字化したデータセットを作成した。抽出した文は Stanford Core NLP[1] によってトークナイズし、結果として約 6402 万語のテキストデータが得られた。

さらに疾患に対応する疾患名の単語を抽出するために、疾患名の辞書を構築した。疾患名辞書は、疾患知識ベース [6] に収録されている英語名を元に構築した。今回用いた疾患は 1,546 件あり、同意語を含めて 1,893 件の疾患名が得られた。

[1] https://stanfordnlp.github.io/CoreNLP

表 1: 疾患知識ベースに収載される疾患英語名フレーズの長さ分布

長さ	1	2	3	4	5	6	7	8	9
頻度	350	945	421	114	46	6	9	1	1

表 1 に示すとおり、疾患名は複数の英単語 (フレーズ) で構築されているものが多い。今回構築した句の長さの平均は 2.28 個であり、最も多い句の長さは 2、その次に 3 であった。疾患について記述されたテキストでは句を考慮して分散表現を構築した方がよいと考えたため、Mikolov ら [4] のフレーズ化の手法により、最大長さ 3 までフレーズ化した。実際の実装としては、word2vec に収録されている word2phrase を用いてフレーズ化を実現した。今回はこのテキストから抽出できる疾患を対象として実験を行った[2]。

2.2 分散表現データの構築

次に、Mikolov ら [4] の Skip-gram モデルを用いて分散表現のモデルを構築した[3]。実装としては word2vec[4] を用い、次元は $N = 200$、ウィンドウ幅 $W = 24$ として、さらに階層的ソフトマックスを用い、イテレート数 15 回で構築した。このパラメータは次元数 $N = 100, 200, 300$、ウィンドウ幅を $W = 4$ から $W = 36$ の間で変更したうえで、後述する性能評価において最高値を示したものである (後述にウィンドウ幅による違いも報告する)。

先に記した医学文献のテキストデータを用いて構築した分散表現データの語彙数は 35,4542 個となり、前述した疾患名辞書を基にして 1,171 疾患分の分散表現ベクトル 1,667 個を抽出することができた。本研究ではこの 1,171 疾患を対象にして、実験を行う。

2.3 疾患名の分散表現を用いた疾患間類似度

本研究では、単語の分散表現を用いた疾患間の類似度スコアとして、以下の定義により計算する。構築した単語 w の分散表現ベクトルを $\mathrm{vec}(w)$ とする。また疾患 D_i の疾患名の集合を $W_n(D_i)$ とする。疾患名の分散表現を用いた疾患間類似度 $sim_{dn}(D_i, D_j)$ は、ベクトル同士のコサイン類似度を $\cos(v_1, v_2)$ としたとき以下のように

$$sim_{dn}(D_1, D_2) = \max_{w_i \in W_n(D_i), w_j \in W_n(D_j)} \cos(\mathrm{vec}(w_i), \mathrm{vec}(w_j))$$

それぞれ疾患名の集合 $W_n(D_i), W_n(D_j)$ についてコサイン類似度が最大のものを類似度として採用した。この類似度を、そのまま疾患間の類似度として出力する。

2.4 評価手法

疾患間類似度の精度は、別途手動生成した疾患類似度データを gold standard とし、生成した類似度データを比較することで評価を行う。疾患知識ベースの中から患者数の多い 80 疾患を抽出し、この 80 疾患のそれぞれに対して、さらに疾患知識ベースの中から 1,408 疾患について疾患全体への類似関係を医師による主観に基づき評価をしている [11]。ただし、医師は、疾患の類似度として、任意の 2 つの疾患が「近い」「遠い」という相対関係を示すことが出来るものの、類似した疾患間でどちらがより近いか、あるいは、どちらがより絶対的に遠いかを示すことが難しい。そこで、疾患類似度データとしては、対象疾患 D_i における比較疾患 D_j に対し 1 から 3 までの相関的な 3 段階でのスコア付けとしデータ化した。具体的には、類似しているほど最も高いスコア 3 を付与し、関係ない疾患には 1 というスコアをつけた。このスコアが高い順に並べたデータを疾患間の類似度データとして用いる。

なお、この gold standard とアルゴリズム出力との比較に際しては、高い類似度であると表示された疾患の予測が重要であり、低い類似度の疾患の予測が評価に過度の影響をうけるべきでない。そこで、ランキング学習などの評価に用いられる Normalized Discounted Cumulative Gain (NDCG) の着想に倣い、評価に勾配を設ける形で開発した Normalized Disease Similarity Measure (NDSM) という指標を用いた [11]。NDSM は、疾患類似度データとアルゴリズム出力の一致度を 0-1 までの実数で表現するもので、1 に近いほど疾患類似度データとアルゴリズム出力の一致度が高い。以下の節では、上位の疾患 250 個までを参照する NDSM($N = 250$) を用いて、我々の発表済み研究 [7, 11] との性能比較を行った。

3 結果と考察

3.1 モデルによる違いの結果

構築したモデルを用いて生成した疾患類似度の NDSM による評価結果を表 2 に示す。また達成性能の評価のため、我々の発表済み研究の性能を、`ICD`、`Prob`、`ICD+Prob`[11] と `ICD+Prob+Loc`[7] として示す。`ICD`

[2] 本稿では予めすべての疾患の英語表現に対してマッチするようにフレーズ化してから抽出していないため、すべての疾患を抽出できていない。すべての疾患名抽出は今後の課題とする。

[3] なお CBOW モデルも検討したものの、事前実験により Skip-gram の方が全体精度が良いとわかったため、本稿では Skip-gram の結果のみを示す。

[4] `https://github.com/tmikolov/word2vec`

表 2: NDSG による評価結果。NDSM で示した結果は抽出できた 73 疾患の平均値となっている。

手法	*NDSM*	最大評価の疾患数
ICD[11]	0.715	-
Prob[11]	0.765	-
ICD+Prob[11]	0.916	17
ICD+Prob+Loc[7]	**0.926**	29
$sim_{dn}(D_i, D_j)$	0.919	27
(合計)		73

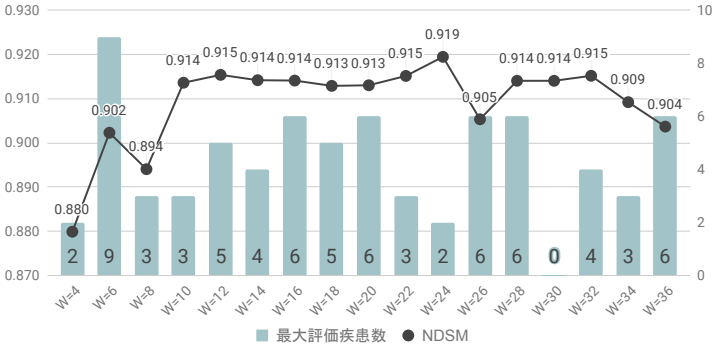

図 1: ウィンドウ幅ごとによる性能の違い。折れ線グラフが NDSM の評価値であり、棒グラフが 73 疾患中で最高評価となったウィンドウ幅の分布を示している。

は疾患の類似度として、疾患が生じるメカニズムを類似度として計算したものであり、Prob は疾患が呈する症状の類似度の類似度を計算したものである。ICD+Prob はこの疾患が生じるメカニズムと疾患が呈する症状の類似度の双方を加味した類似度計算手法である。

疾患が生じるメカニズムとは、たとえば、「細菌感染により生じる」とか「免疫不全により生じる」といったもので、ICD と呼ばれる疾患の分類体系 [9] を利用することにより定量化できる。「疾患が呈する症状」による類似度とは、たとえば、「腹痛を生じる疾患同士」をより類似性が高いとみなし定量化するものである。

我々の研究により、この「疾患の生じるメカニズムによる類似度」(ICD) と「疾患の呈する症状による類似度」(Prob) を機械学習により組み合わせることで、高い性能が得られることが示された [8]。これは、医師が異なった根拠を疾患毎に使い分けながら類似度を判断することを示唆する。

この観察を元に、さらなる性能向上に取り組んだのが文献 ICD+Prob+Loc[7] となる。これは ICD+Prob に加えて、疾患が影響を与える臓器や場所等の特徴 (+Loc) を取り入れたモデルであり、ICD+Prob 以上の性能が得られることが明らかとなった。

表 2 に各手法の結果の平均値を示す。$sim_{dn}(D_i, D_j)$ は今回計算した疾患名の分散表現に基づく類似度手法となる。なお、文献 [7, 11] では 80 疾患を対象にしているが、本稿では、提案手法が利用した辞書において 7 疾患ほどの抽出に困難が生じたため、手法間での評価を可能とするために全てのセッティングにおいて 73 疾患を対象とした評価としてある。また同様に、比較対象の疾患としても、1,408 疾患ではなく、1,171 疾患を対象とした結果となっている。表 2 中の「最大評価の疾患数」ではさらに、3 手法中の中で最大の評価だった疾患数を示している。分散表現を用いた手法ではとくに「境界型人格障害」「カンジダ症」「子宮内膜症」「アトピー性皮膚炎」といった疾患 27 件は最高精度であることが分かった。しかし、分散表現を用いた手法は、ICD や Prob また ICD+Prob よりは性能が良いものの、必ずしもすべての疾患において最高性能を示したわけではないことが分かる。また ICD+Prob+Loc との比較において性能が低いことが分かった。

3.2 文脈幅による違いの結果

ベクトル次元数を 200 固定とし、ウィンドウ幅 $W = 4, 6, 8, ..., 36$ としてパラメータを変更した上での結果を比較した。図 1 にウィンドウ幅ごとによる性能の違いを示す。図 1 の折れ線グラフで示している箇所が NDSM での評価値の平均となる。$W = 24$ のときの NDSM が 0.919 となり最高精度となることが分かった。

この最高値は全 73 疾患それぞれの評価値の平均であり、NDSM の平均値が高くとも、各疾患の評価値は、ウィンドウ幅等の変化によって、0-1 の間で上昇、下降する。そこで、ウィンドウ幅を $W = 4$ から $W = 36$ へと推移させつつ、構築された分散表現による類似度計算結果について、各疾患毎に、NDSM 上の最高値を示したウィンドウ幅を記録した。その上で、各ウィンドウ幅毎に、最高値を示した疾患の数をカウントした。これらの結果を図 1 に棒グラフにて示す。平均値が最高値を示した $W = 24$ において、最高値を示した疾患は 2 件のみであり、$W = 6$ のときで 9 件となった。図 1 のグラフをみて分かるように、73 疾患の中で最高精度となるウィンドウ幅の分散表現にはばらつきがあり、このことから、ウィンドウ幅に比して平均性能が単純上昇するのではなく、各疾患における疾患類似度計算の性能が独立して変動していることが示唆される。

3.3 考察

Skip-gram のようなモデルで構築された分散表現では、文脈が類似しているような疾患同士を類似関係とするようなモデルを構築しベクトルを生成する。今回の実験によって、こうした分散表現を用いることで、低コストに疾患間の類似度を生成し、これがある程度の妥当性を有していることを実証した。とりわけ、「境界型人格障害」「カンジダ症」「子宮内膜症」「アトピー性皮膚炎」などの 27 疾患を対象とした場合は、既存手法の性能を上回ることが分かった。分散表現によって生成されたベクトルにおいては、類似した発症メカニズムの疾患同士が類似したり、疾患が呈する症状に類似性がみられたりと、医学文献の文脈に現れる疾患の対象部位の同一性などを手がかりとして、類縁疾患を算出しているものと考えられる。

一方で、医学文献から生成したモデルに基づく分散表現では、より細かな配慮を行った類似度計算 ICD+Prob+Loc[7] に性能面で差が生じることも明らかとなった。我々の研究では、医師が認知する疾患の類似度には、疾患の発生するメカニズム (ICD)、疾患の呈する症状 (Prob) に加えて、疾患が影響を与える解剖学的な部位 (Loc) といったいくつかの要因があり、医師はそのそれぞれを状況に応じて使い分けていることが示唆されている。分散表現は、この使い分けを機械学習により近似したアルゴリズム (ICD+Prob+Loc) に劣ることが示された。

今回、疾患類似度の背景にあるこうしたミクロなメカニズムを捨て去り、文脈的な情報を分散表現によりベクトル化した。その情報中には、疾患のメカニズム、症状、疾患の発生部位が混在している可能性が高いが、各情報の精度が担保されていない。また、医師が行っているような各類似度の使い分けを正しく反映しているかが定かでない。したがって、分散表現の疾患類似度計算への活用に際して、疾患概念のダイレクトなベクトル化による性能向上には限界が考えられる。今後適切な各情報の精度管理の環境整備のうえで、医学文献からの各種類の類似度情報の抽出に取り組むことにより、さらなる性能向上がもたらされる可能性がある。

4 おわりに

本研究では Skip-gram モデルによって疾患名の分散表現データを構築し、疾患名の分散表現ベクトルを基に疾患間類似度計算を行い、疾患類似度計算における有効性を検証した。その結果、疾患名を対象とした分散表現によって、疾患類似度を低コストに一定の性能で計算しうることが明らかとなった。一方で、個別の疾患においては本手法が高性能を示すケースがあったものの、全体平均でみる限り、先行研究の最高精度には達しないことが明らかとなった。

次のステップとして、疾患の類似度に関連していると考えられるいくつかの要因を精査し、そのそれぞれの要因に対して分散表現がいかに貢献しうるか、より細かな検討が望まれる。とりわけ、従来手法においては、疾患の呈する症状や影響を及ぼす解剖学的部位のデータ生成に多くのコストを要し、網羅性に難が生じていた。分散表現が、疾患の位置情報を低コストに、また、網羅的に生成しうるのであれば、従来手法に組み込むことによりさらなる高精度化が実現する可能性がある。

参考文献

[1] Lance De Vine, Guido Zuccon, Bevan Koopman, Laurianne Sitbon, and Peter Bruza. Medical semantic similarity with a neural language model. In *Proceedings of the CIKM 2014*, pp. 1819–1822, 2014.

[2] Saurav Ghosh, Prithwish Chakraborty, Emily Cohn, John S. Brownstein, and Naren Ramakrishnan. Characterizing Diseases from Unstructured Text: A Vocabulary Driven Word2vec Approach. In *Proceedings of the CIKM 2016*, pp. 1129–1138, 2016.

[3] Sachin Mathur and Deendayal Dinakarpandian. Finding disease similarity based on implicit semantic similarity. *Journal of Biomedical Informatics*, Vol. 45, No. 2, pp. 363–371, 2012.

[4] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of the NIPS 2013*, pp. 3111–3119, 2013.

[5] Takashi Okumura and Masaki Tagawa. Hierarchical representation of differential diagnosis lists for clinical decision support systems. In *Proceedings of the PervasiveHealth '14*, pp. 231–234, 2014.

[6] Takashi Okumura, Hiroaki Tanaka, Mai Omura, Maori Ito, Shin'ichi Nakagawa, and Yuka Tateisi. Cost decisions in the development of disease knowledge base: A case study. In *Proceedings of the IEEE BIBM 2014*, pp. 62–69, 2014.

[7] Mai Omura, Noboru Sonehara, and Takashi Okumura. Practical approach for disease similarity calculation based on disease phenotype, etiology, and locational clues in disease names. In *Proceedings of the IEEE BIBM 2016*, pp. 1002–1009, 2016.

[8] Mai Omura, Yuka Tateishi, and Takashi Okumura. Disease similarity calculation on simplified disease knowledge base for clinical decision support systems. In *In Proceedings of the FLAIRS Conference 2015*, pp. 501–506, 2015.

[9] World Health Organization. *ICD-10: International statistical classification of diseases and related health problems.* World Health Organization, 2011.

[10] 岡崎直観. 言語処理における分散表現学習のフロンティア (<特集>ニューラルネットワーク研究のフロンティア). 人工知能学会誌, Vol. 31, No. 2, pp. 189–201, 2016.

[11] 大村舞, 建石由佳, 奥村貴史. 簡易疾患知識ベースを基にした疾患間の類似度計算. 第 103 回知識ベースシステム研究会予稿集, pp. 42–47, 2014.