オントロジーの活用によるフレーム意味論の単語重み付けへの適用事例

An application of the frame semantics to term weighting by using ontologies

松尾亮輔[1*]　Ho Tu Bao[1]
Ryosuke Matsuo[1]　Tu Bao Ho[1]

[1] 北陸先端科学技術大学院大学
[1] Japan Advanced Institute of Science and Technology

Abstract: This paper proposes a framework of the application of the frame semantics to term weighting and illustrate a semantic term weighting method based on the patients' severity as one example of the application. The frame semantics enables to consider various term's semantics based on the construal and exploits world knowledge such as encyclopedic knowledge to the semantic understanding of terms for computers. The framework to apply the frame semantics to term weighting is effective as the patients' severity-based term weighting brought about high performance in mortality prediction. Ontologies are useful for mapping terms to encyclopedic knowledge of a causes of death ranking by exploiting the terms' concepts to capture the terms' weights regarding the patients' severity.

1　はじめに

フレームの概念を用いて状況を特徴づけるあるフレームを考えることで，そのフレームがある単語の意味を割り当てるとするのがフレーム意味論である [1]．フレーム意味論は人間の認知プロセスに着目している．単語の意味を概念的に考えると，主体がどのように単語の意味を捉えているか [2] が認知言語学において重要であり，その捉え方にあたるフレームを用いることで語の意味は異なる捉え方を可能にする点がフレーム意味論の特徴である．また，フレームとは背景知識で，それは多くの場合人々の日常生活を通じて形成された経験的知識 [3, 4] である，またはそれと似た意味となる人間が様々な経験を通して身に付けた百科事典的な知識 [5] のことであるため，語の意味を理解するのに百科事典的知識を参照していることもフレーム意味論の特徴である．語は百科事典的意味とそれとは対称的な基本的意味を有している [5]．ここで百科事典的意味とは,「ある語が指し示す対象（の典型的なもの，代表的なもの）がもつもろもろの性質・特徴，さらには，その対象と関連をもつ（たとえば，その対象から連想される）様々な事柄」 [6] で，基本的意味とは,「語が指し示す対象のすべてに該当する意味であり，かつ，類義語との弁別的特徴を含むもの」と定義されている [5]．

本論文では，捉え方によって異なる意味が考えられ，語の意味を理解するのに百科事典的知識を参照するフレーム意味論の考え方を単語の重み付けという計算手法に適用する．単語の重み付けは単語の重要度という観点から単語に対して重みを数値で与える手法である．ある単語が重要かそうでないかは，フレーム意味論の考え方と同様に人間の認知プロセスを考慮すると，同じ語であっても主体の視点の違いによって重要度の程度は変化し得る．例えば，医学単語における癌を例に挙げると，癌の基本的意味は悪性の腫瘍 [7] であり，百科事典的意味は，例えば，死因ランキング 2 位に該当する危険な疾病である，という患者状態の重症度という視点や，手術の複雑度，あるいは医療費の程度といった様々な視点による意味が考えられる．コンピュータがこれらの視点から単語の意味を捉えるには百科事典的知識を組み込むことが求められるが，単語の重要度の決定の仕方は一意に決まらず多様であることが人間の認知プロセスの観点から単語の重み付けを照射することで見えてくる．

本研究では，状況を特徴づけるフレームが単語の意味を割り当てるとするフレーム意味論の考え方を単語の重み付け手法へ適用するためのフレームワークとその適用事例を報告する．また，コンピュータの単語の意味理解において，オントロジーの活用が百科事典的

*連絡先：北陸先端科学技術大学院大学知識科学研究科
〒 923-1292 石川県能美市旭台 1-1
E-mail: matsuor@jaist.ac.jp

知識のような世界知識へ単語を結びつけるのに有効であることも併せて報告する．

2　提案方法

2.1　フレームワーク

フレーム意味論の考え方を単語の重み付けへ適用するためのフレームワークは，以下の 3 つのステップで構成され，図 1 で表される．

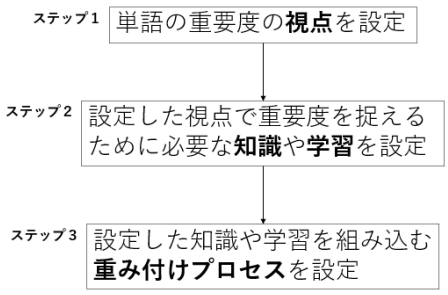

図 1: フレームワーク

最初のステップでは，単語の意味を考慮しながら重み付けをするにあたり，どのような視点からその単語の重要度を捉えるかを設定する．その際，その視点による重み付けをどのような文書分析に適用するかという目的も併せて考慮する．次に，その設定した視点からコンピュータが単語の重要度を捉えるために必要な知識や学習を設定する．最後に，その設定した知識や学習を組み込む重み付けプロセスを設定する．これら 3 つのステップによりフレーム意味論の考え方を適用した単語の重み付けが実現される．

2.2　事例

本論文では，フレーム意味論の考え方を重み付け手法に適用した事例として，我々が提案している電子カルテの医学文書を用いた患者の重症度という視点による重み付け手法 [8] を挙げる．

まず，患者の重症度を単語の重要度の視点として設定し，死亡予測やリスク予測といった目的に適した重み付けを開発することとする．次に，患者の重症度を考慮した重みをコンピュータが捉えるために死因ランキング [9] という百科事典的知識を設定する．なぜなら，そのランキングの持つ線形関係に基づいて患者の重症度を数値で単語に付与できるからである．ここで問題は単語をどのようにしてその知識に結びつけるかであるが，死因ランキング内の単語は ICD-10 [10] のコードと対応していることから，ICD-10 の医学知識を用いることとする．電子カルテ内の単語に ICD-10 コードを付与するために，医学のオントロジーを活用する．特に，単語の概念情報に基づいて医学単語を識別する辞書的役割，及びその概念情報を用いて ICD-10 の知識へマッピングするために活用する．したがって，ここで用いる知識は，死因ランキング，ICD-10 及び，医学オントロジーである UMLS [11] と BioPortal [12] である．最後に，このような知識を組み込む重み付けのプロセスを以下のように設定する．

1. 電子カルテ内の単語のタイプの識別
2. 医学重要度の決定

上記の 2 つのステップにより設定した知識を組み込み，重症度を考慮した重み付け手法を開発する．以下にそれぞれのステップの詳細を述べる．

はじめのステップでは，電子カルテ内の単語を非医学単語，医学単語，ICD-10 単語の 3 つのタイプに分ける．まず，UMLS [11] という医学オントロジーにより単語の概念に基づいて医学単語を識別する．その際，電子カルテの文書内の単語を UMLS の概念である UMLS concept へマッピングするために MetaMap [13] を用いる．ここでは，オントロジーを辞書的役割として用いることで，同じ意味であっても異なる表記がされている同義語といった問題に対応しながら医学単語を識別する．次に，医学単語の場合は UMLS から得られる UMLS concept の概念情報を用いて，BioPortal [12] 上で医学単語の ICD-10 コードの取得を試みる．そして，ICD-10 コードを持つ医学単語を ICD-10 単語とする．ここでは，オントロジーを単語の概念情報を活用しながら，単語を他の医学知識へマッピングするために活用している．図 2 はこのようなオントロジーを活用して電子カルテ内の単語を 3 タイプに区分するプロセスを示している．

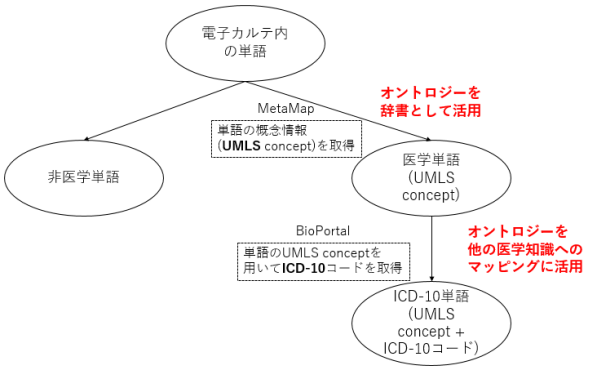

図 2: オントロジーの活用による単語タイプの識別

次のステップでは，オントロジーの活用により識別した ICD-10 単語が持つ ICD-10 コードを用いて，単

語を百科事典的知識である死因ランキング [9] に結びつけることで，医学重要度という患者の重症度を考慮した重みを捉える．ランキングの線形関係を保持しながら，ランキングに対応する ICD-10 コードに基づき，ICD-10 単語に対して医学重要度を付与する．以下の表 1 はトップ 15 の死因名と対応する ICD-10 コード及び医学重要度の値を示している．ここでは，以下の式 (1) を用いて医学重要度を求めている．

表 1: トップ 15 の死因ランキング [9] の死因名とそれに対応する ICD-10 コード及び医学重要度の値

ランク	死因名	ICD-10 コード	値
1	Disease of heart	I00-I09, I11, I13, I20-I51	0.7
2	Malignant neoplasms	C00-C97	0.66
3	Chronic lower respiratory diseases	J40-J47	0.62
4	Cerebrovascular diseases	I60-I69	0.58
5	Accidents (unintentional injuries)	V01-X59, Y85-Y86	0.54
6	Alzheimer's disease	G30	0.49
7	Diabetes mellitus	E10-E14	0.45
8	Nephritis, nephritic syndrome and nephrosis	N00-N07, N17-N19, N25-N27	0.41
9	Influenza and pneumonia	J09-J18	0.37
10	Intentional self-harm (suicide)	U03, X60-X84, Y87.0	0.33
11	Septicemia	A40-A41	0.29
12	Chronic liver disease and cirrhosis	K70, K73-K74	0.25
13	Essential hypertension and hypertensive renal disease	I10, I12, I15	0.21
14	Parkinson's disease	G20-G21	0.16
15	Pneumonitis due to solids and liquids	J69	0.12

$$v_i = \frac{(v_{max} - v_{min}) \times (\xi - i + 1)}{\xi} \quad (1)$$

v_i は順位 i 番目のランクに対応する医学重要度である．v_{max} と v_{min} は医学重要度の最大値と最小値にあたり，それぞれ 0.9，0.2 としている．ξ はランクにより区分けされるグループの総数である．ただし，死因ランキングに該当しない ICD-10 単語を 16 番目のグループ，ICD-10 単語ではない医学単語を 17 番目のグループとし，それぞれの医学重要度は 0.08，0.04 としている．よって，ξ の値は最終的に 17 となっている．

ここで求めた重症度を考慮した医学重要度の重み (MED) は，単語の出現頻度に基づいて重要度が計算される単語の重み付けにおいて伝統的な手法である TFIDF [14] と以下のように組み合わされる．

$$Proposed\ weight = \alpha^1 * TFIDF * \{1 + (\alpha^2 * MED)\} \quad (2)$$

α^1 と α^2 は TFIDF と MED の重みの係数で，それぞれ 0.5，1.5 としており，TFIDF の影響を小さくし，医学重要度の影響を大きくするためにこれらの係数を用いている．

死因ランキングの活用による医学重要度を考慮した重み付け手法の有効性を検証するため，電子カルテを含む MIMIC II [15] のデータベースを用いた死亡予測の実験が行われ，その結果は以下の図 3 [8] である．ここでは，ベースラインとなる TFIDF と他の 3 手法を比較している．TFIDF-MED-Ranking と表記されている手法が，死因ランキングを活用して患者の重症度を捉えた手法である．TFIDF-MED-1 と表記されている手法は，医学単語に対して医学重要度を一定で付与した場合である．TFIDF-MED-2 と表記されている手法は，ICD-10 コードを持たない医学単語よりも ICD-10 単語により高い重みを与えるという 2 つのパターンで医学重要度を捉えた場合である．そして，Support vector machine（RBF カーネル）を用いて 5-fold cross-validation を 500 回実施した累積平均による精度を手法ごとで示している．結果から，TFIDF-MED-Ranking が最も高い精度であるため，フレーム意味論の考え方を適用し，百科事典的知識である死因ランキングを活用している重み付け手法は，死亡予測というタスクにおいて適した手法であることがいえる．

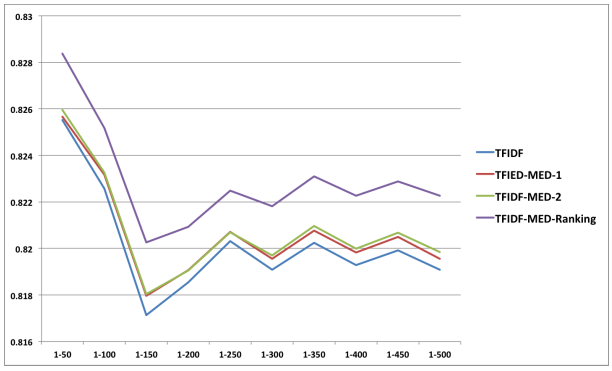

図 3: 累積平均による精度の推移グラフ

3 むすび

本論文ではフレーム意味論の考え方を単語の重み付けという計算手法に適用するためのフレームワークとその適用事例を述べた．フレーム意味論の考え方を用いることで，単語の意味を理解する際，人間の認知プロセスと同じように語の意味は様々な視点で捉えられるという前提を保持し，かつ百科事典的知識を参照した意味理解が可能になった．本論文ではこのような特徴を取り入れた単語の重み付けのフレームワークを提案し，それを適用した例として，電子カルテを用いた患者の重症度を考慮した重み付け手法を挙げた．フレーム意味論の考え方を取り入れた重症度を考慮した重み

付け手法は死亡予測の実験において有効性が示されていることから，提案するフレームワークを用いることで，ある目的に適した重み付け手法の開発が可能になることを示した．そのフレームワークにおいて，特にオントロジーが患者の重症度という観点からコンピュータが単語の意味を捉えるため，死因ランキングという百科事典的知識へ単語を結びつけるのに有効であることを示した．

単語の重み付けは，文書をベクトル空間上で計算可能な形式へ変換できる文書分析において基礎となる手法であることから，フレーム意味論の考え方を取り入れることで，多様な視点の意味を保持しながら，様々な文書分析への適用が可能になると考えられる．

参考文献

[1] Fillmore, C.: Frame semantics, *Linguistics in the morning calm*, pp. 111–137 (1982)

[2] Langacker, R. W.: Foundations of cognitive grammar: Theoretical prerequisites, *Stanford university press*, Vol. 1 (1987)

[3] 藤井聖子, 小原京子: 構文研究の理論と実践 (6) フレーム意味論とフレームネット, 英語青年, Vol. 149, No. 6, pp. 373–376 (2003)

[4] 小原京子: フレーム意味論と日本語フレームネット (特集 文の意味と語の意味), 日本語学, Vol. 25, No. 6, pp. 40–52 (2006)

[5] 籾山洋介: メタファーの認知的基盤と経験的基盤, 言語文化研究叢書, Vol. 7, pp. 97–111 (2008)

[6] 籾山洋介: 1-8. 認知言語学, 言語科学の百科事典, pp. 157–177 (2006)

[7] 三省堂 Web Dictionary `http://www.sanseido.net/` (2016/10/31 アクセス)

[8] 松尾亮輔, Ho Tu Bao: 重症度を考慮した医学単語重み付け手法による死亡予測, *2016* 年度人工知能学会全国大会（第 *30* 回）, 4D1-4in2 (2016)

[9] Murphy, S. L., Xu, J., Kochanek, K. D.: Deaths: final data for 2010, *National vital statistics reports: from the Centers for Disease Control and Prevention, National Center for Health Statistics, National Vital Statistics System*, Vol. 61, No. 4, pp. 1–117 (2013)

[10] World Health Organization.: International statistical classification of diseases and related health problems, *World Health Organization*, Vol. 1 (2004)

[11] Bodenreider, O.: The unified medical language system (UMLS): integrating biomedical terminology, *Nucleic acids research*, Vol. 32, No. suppl 1, pp. D267–D270 (2004)

[12] Noy, N. F., Shah, N. H., Whetzel, P. L., Dai, B., Dorf, M., Griffith, N., et al.: BioPortal: ontologies and integrated data resources at the click of a mouse, *Nucleic acids research*, Vol. 37, pp. W170–W173 (2009)

[13] Aronson, A. R.: Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program, *Proceedings of the AMIA Symposium*, pp. 17–21 (2001)

[14] Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval, *Information processing & management*, Vol. 24, No. 5, pp. 513–523 (1988)

[15] Saeed, M., Villarroel, M., Reisner, A. T., Clifford, G., Lehman, L. W., Moody, G., et al.: Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II): a public-access intensive care unit database, *Critical care medicine*, Vol. 39, No. 5, pp. 952–960 (2011)