一般口演 | 知識工学
一般口演10
知識工学
2019年11月23日(土) 09:00 〜 11:00 C会場 (国際会議場 2階国際会議室)

[3-C-1-03] 医学用語抽出のための文字列類似度

○篠原 恵美子[1]、河添 悦昌[1]、今井 健[2]、大江 和彦[3] （1. 東京大学大学院医学系研究科医療AI開発学講座, 2. 東京大学大学院医学系研究科疾患生命工学センター, 3. 東京大学大学院医学系研究科社会医学専攻医療情報学分野）

キーワード：Natural Language Processing, String Similarity Metric, Medical Informatics Computing

【背景・目的】診療録の自由記述を利活用する際には、そこに現れる医学用語の抽出が有用である。しかし自由記述は表記揺れを含むため、対応する用語が定義されているはずの医学用語集であっても、完全一致検索ではその用語が得られない場合がある。また、医学用語集として利用可能なリソースは管理者や目的がそれぞれ異なるため、カバーする異表記のバリエーションがリソースによってさまざまであり、用語集側で自然言語処理を目的として表記揺れをすべて吸収するのは現実的ではない。
この課題の解決策として類似文字列検索が挙げられるが、編集距離では文字の意味を考慮できず、機械学習では教師データの用意が困難である。本研究ではこれらの問題点を克服する文字列類似度指標を提案する。

【方法・結果】提案する文字列類似度は、文字を2段階で正規化し、その結果を用いるものである。正規化はユニコード正規化を拡張したもので、文字ごとに定義した変換規則を再帰的に適用する。変換規則はユニコード正規化で用いられるものに異体字や長音記号などの規則を追加し、ユニコード正規化の canonical formに近い規則集合を適用したものを第一段階、全規則を適用したものを第二段階とする。2つの文字列の類似度は、まず第二段階の結果を比較して合致しなければ0とし、合致した場合は第一段階の結果を比較し編集距離を拡張した文字列類似度を用いるものである。
評価として、標準病名マスターの索引テーブルの異字体区分が1または2（誤字/異字）、かつかな漢字区分が1（漢字文字列）の索引用語に対し、他の索引用語から最も類似度の高い用語を検索したとき、その対応用語コードが同一であるかを調査した。その結果、4384件のうち4374件に対し同一対応用語コードが得られ、誤りの原因は9件が変換規則の不足、2件がマスターに定義された用語の曖昧性であった。

医学用語抽出のための文字列類似度

篠原恵美子[*1]、河添悦昌[*1]、今井健[*2]、大江和彦[*3]
*1 東京大学大学院医学系研究科医療 AI 開発学講座、*2 東京大学大学院医学系研究科疾患生命工学センター、
*3 東京大学大学院医学系研究科社会医学専攻医療情報学分野

String similarity metric for clinical concept extraction

Emiko Shinohara[*1], Yoshimasa Kawazoe[*1], Takeshi Imai[*2], Kazuhiko Ohe[*3]
*1 Artificial Intelligence in Healthcare, Graduate School of Medicine, The University of Tokyo,
*2 Center for Disease Biology and Integrative Medicine, Graduate School of Medicine, The University of Tokyo,
*3 Department of Biomedical Informatics, Graduate School of Medicine, The University of Tokyo

While clinical concept extraction is helpful to utilize clinical notes, simple string match is not enough as spelling variation exists. It is not realistic to claim each terminology resources, that are managed by different groups for different purposes, should cover all spelling variations for natural language processing. Soft match is a solution for the problem, however, edit distance does not consider semantics of characters and machine learning is difficult at the point of preparing training data.
The study proposes a string similarity metric to solve the problem. It utilizes the result of two step character normalization. The normalization is recursive application of conversion rules defined for each character. The rules are based on Unicode normalization, adding some rules including alternative forms of character. The first step is the result of normalization, which almost the same as canonical form of Unicode normalization, using subset of the all rules and the second is the result of all the rules. The similarity of two strings is 0 if the second result is different, else the edit distance like metric of the first result. The proposed method could identify the correct variation term of 99.7% (4369/4384) in the Japanese Standard Disease Code Master.

Keywords: Natural Language Processing, String Similarity Metric, Medical Informatics Computing

1. 緒論

診療録の自由記述を利活用するには自然言語処理が必要であるが、その際には医学用語の抽出が有用である。しかし自由記述は表記揺れを含むため、対照先として用いる用語集に対応する用語が定義されていても、完全一致検索ではその用語が得られない場合がある。また、医学用語集として利用可能なリソースは管理者や目的がそれぞれ異なるため、カバーする異表記のバリエーションがリソースによってさまざまであり、用語集側で自然言語処理を目的として表記揺れをすべて吸収するのは現実的ではない。

この課題の解決策として、前処理として行う文字の正規化が挙げられる。よく知られているものとしてはユニコード正規化があるが、大文字と小文字が統一されないなど不十分な場合がある。しかし大文字小文字を区別したほうがよい場合もあり、正規化の規則の最適解は存在しないと言える。

他の方法として類似文字列検索がある。ここでは文字列間の類似度尺度が必要である。編集距離では文字の意味を考慮せず、例えば「一型糖尿病」に対して「1 型糖尿病」と「2 型糖尿病」が同じ距離となる。一方、機械学習ではこのような問題を回避できると考えられるが、教師データの用意が困難である。Aramaki らは同じ用語ならば英訳が同一であることを利用して複数の医学用語集から教師データを自動生成した[1]が、用語集に収載される語と診療録で用いられる表現とでは表記ゆれの性質が異なると考えられる。また、類似度は一般的にゼロになることが無く、用いる用語集に対応する用語が含まれていなかった場合にも「最も似ている用語」が得られてしまう。これを避けるために閾値を設定するが、適切な値を決めるのは難しい。また、用語集に収載される全用語と類似度を計算するのは計算時間がかかる。高速な類似文字列検索アルゴリズムは存在する[2]ものの、文章の全部分文字列に対して適用するのは非現実的である。

2. 目的

本研究ではこれらの問題点を克服する文字列類似度指標を提案する。すなわち、文字の意味を考慮し、教師データを必要とせず、類似度として 0 を出力可能で、類似度計算対象となる候補用語集合のサイズが小さいような文字列類似度指標である。

3. 方法

3.1 提案手法

提案する文字列類似度は、文字を 2 段階で正規化し、一方で候補用語集合の絞り込みを、もう一方で類似度計算を行うものである。以下では文字の正規化と類似度算出のアルゴリズムを述べる。

3.1.1 文字の正規化

まず提案手法のベースとなるユニコード正規化のアルゴリズム[3]の概要について述べる。各文字に対し、それが正規化可能であれば ”decomposition rule” と呼ばれる 1 文字以上の文字列への対応関係が変換規則として付与されている。この変換規則を再帰的に適用し、得られた文字列を各文字に付与されている数値の大小関係を元に並べ替えを行う。変換規則のうちあるサブセットのみを適用した場合が canonical decomposition、全てを適用した場合が compatibility decomposition と、2 段階の正規化が定義されている。前者は元の文字列と等価であり、後者はいくらかの情報が欠落したものである。

緒論で述べたとおり、ユニコード正規化では大文字小文字の統一ができないなど辞書検索に不十分である。そこで本研究の提案手法ではユニコード正規化に変換規則を追加したものを文字の正規化手法として用いる。本稿では変換規則は、意味を変えないものとして、文字情報基盤事業[4]の縮退マッ

プ、日本医学会 医学用語辞典の漢字表記規則 [5]、web から人手で収集した旧字体のほか、長音やハイフンなどをまとめる規則、制御文字と異体字セレクタを削除する規則を追加した。また、意味を変えうる規則として、意味が似ている文字への変換規則、形が似ている文字への変換規則を追加した。

以上の変換規則のうち一部を適用したものを第 1 段階、全てを適用したものを第 2 段階の正規化結果とする。前者は類似度計算のため、後者は候補用語集合の絞り込みのために用いる。第 1 段階に該当する変換規則は、ユニコードで定義される変換規則のうち canonical decomposition および compatibility formatting tag が<font>、<noBreak>、<wide>、<narrow>のいずれかであるもの、そして追加規則のうち意味を変えないものとした。

表 1 追加した変換規則

	説明	例
<variant>	意味を変えない規則	頸→頚
<semSim>	意味が似ている文字への変換規則	あ→ア
<visSim>	形が似ている文字への変換規則	読点→カンマ 波線→ハイフン

3.1.2 文字列類似度

2 つの文字列の類似度は、まず第 2 段階の結果を比較して合致しなければ 0 とする。つまり、候補用語集合は第 2 段階の結果が合致したもののみとなる。合致した場合は第 1 段階以降の正規化過程を比較し編集距離を拡張した尺度を文字列類似度として算出する。拡張の内容は操作コストの設定であり、挿入・削除のコストを置換コストの理論的上限値+1、置換のコストを以下に述べる文字の距離とした。

正規化された 2 つの文字の距離は、第 1 段階から第 2 段階までの間に適用された変換規則のカテゴリのうち 2 文字間で重複しないものの種類数とした。

3.2. 評価

予備的な評価として、ICD10 対応標準病名マスターを用いた実験を行った。このタスクは難易度が低く、他の手法との比較ではなく、最低限解けるべきものを本当に解けるかを確認するという意図での実験設定である。

3.2.1 方法

ICD10 対応標準病名マスターver.5.00 の索引テーブルの異字体区分が 1 または 2(誤字/異字)、かつ、かな漢字区分が 1(漢字文字列)の索引用語をクエリとして、それ以外の索引用語から最も類似度の高い用語を検索したとき、その対応用語コードが同一であるかを調査した。また、誤り分析を行い、改善点の洗い出しを行った。

3.2.2 結果

4384 クエリのうち 15 件について検索結果に異なる対応用語コードを持つ用語が含まれていた。表 2 に誤り分析の結果を示す。8 件は正解と異なる用語コードが得られ(表中「FP」)、7 件は検索結果が 0 件であった(表中「FN」)。15 件中 4 件はマスターに定義された用語の曖昧性による誤りで、11 件が提案手法による誤りであった。11 件中 9 件は変換規則に起因するもので、うち 8 件はクエリと正解用語の差がかな表記・漢字表記であった。他の変換規則の不足としては、クエリ「先天性第 11 因子欠乏症」に対し「先天性第 XI 因子欠乏症」が検索できなかった例があり、これは 11 と XI が正規化第 2 段階で同じ文字列にならなかったためであった。また、文字列類似度による誤りは 2 件で、クエリ「先天性第 2 因子欠乏症」に対し正解「先天性第Ⅱ因子欠乏症」と誤り「先天性第 11 因子欠乏症」の両方を最も類似しているとしたものであった。これは文字間の距離の定義を変換規則のカテゴリ数としたことに起因する。

表 2 誤り分析の結果

	マスター	変換規則	類似度
FP	4	2	2
FN		7	

4. 考察

従来の文字の正規化は結果を唯一に固定しており、正規化規則の追加には precision と recall のトレードオフが発生していた。機械学習はそれを回避しうるものの、学習用データや計算時間の点で課題があった。本研究で提案した手法はこれらの課題をシンプルに解決するものと考える。

前章で示した評価の結果、提案手法の 3 つの課題が明らかとなった。1 つはかな-漢字の表記ゆれに対応できていないという点である。これに対応するためには変換規則に漢字→かな表記の規則、またはその逆を追加することが必要となるが、読み/漢字表記の対応関係は文脈依存であり、文字に対して画一的な変換規則では不十分である。文脈を考慮するか、または 1 文字に対して複数の異なる変換規則を適用可能とするようなアルゴリズムおよび類似度定義の変更が必要と考えられる。

2 つめの課題は文字列間の表記ゆれである。提案手法では文字に対する変換規則を用いるため、文字列間関係である「11→XI」(またはその逆)を表現できない。このような変換規則を導入しようとすると先述の課題と同様に文脈依存の問題が出てくる。したがって提案手法の改善案としては 1 つめの課題と同じである。

3 つめの課題は文字列類似度の定義、あるいは変換規則の管理に関するものである。文字間の距離として適用された変換規則のカテゴリの種類数を用いたが、同一カテゴリ内の複数の規則が適用されることがあり、それが距離に反映されていなかったことが誤りの原因であった。変換規則のグルーピングを適切に行う必要があると考えられる。

5. 結論

文字の意味を考慮し、教師データを必要とせず、類似度として 0 を出力可能で、類似度計算対象となる候補用語集合のサイズが小さいような文字列類似度指標を提案し、予備評価により改善点の洗い出しを行った。今後は症例報告や診療録での評価を行う予定である。

参考文献

1) Eiji Aramaki, Takeshi Imai, Kengo Miyo, Kazuhiko Ohe: Orthographic Disambiguation Incorporating Transliterated Probability. IJCNLP 2008, pp. 48-55.
2) 岡崎直観, 辻井潤一. 高速な類似文字列検索アルゴリズム. 情報処理学会創立 50 周年記念全国大会, 2010; 567-9.
3) The Unicode® Standard Version 12.0 - Core Specification. The Unicode Consortium, Mountain View, California, USA. [https://www.unicode.org/versions/Unicode12.0.0/UnicodeStandard-12.0.pdf (cited 2019-Aug-16)].
4) 文字情報基盤事業. 独立行政法人 情報処理推進機構. [https://mojikiban.ipa.go.jp/ (cited 2019-Aug-16)]
5) 日本医学会 医学用語辞典 WEB 版 付表 1 漢字表記のゆれ [http://jams.med.or.jp/dic/kanji_variance2.html (cited

2019-Aug-16)]