社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

N-gramに基づく用例対訳検索手法

田淵　裕章† 　坂本　廣†† 　北村　泰彦††

† 関西学院大学大学院理工学研究科　〒669–1337 兵庫県三田市学園2丁目1番地
†† 関西学院大学理工学部　〒669–1337 兵庫県三田市学園2丁目1番地
E-mail: †tabuchi@ksc.kwansei.ac.jp, ††{hiroshi0606,ykitamura}@kwansei.ac.jp

あらまし　外国人を対象とした多言語医療通訳支援を目的として多数の医療用例対訳が収集されている．用例対訳を検索する場合に，従来のキーワード検索手法は，検索の絞り込みや用例表現のゆらぎに関して問題がある．そこで本研究ではN-gramを用いた類似度計算に基づく用例検索手法を提案し，その有効性を示す．本手法は，入力された検索文が新たな用例として活用できるという利点も備えている．

キーワード　類似文章検索，N-gram，情報検索，用例対訳

Search Method for Parallel Texts Using N-gram

Hiroaki TABUCHI†, Hiroshi SAKAMOTO††, and Yasuhiko KITAMURA††

† Graduate School of Engineering, Kwansei Gakuin University gakuen 2–1, Sanda-shi, 669–1337, Hyougo, Japan
†† Faculty of Engineering, Kwansei Gakuin University gakuen 2–1, Sanda-shi, 669–1337, Hyougo, Japan
E-mail: †tabuchi@ksc.kwansei.ac.jp, ††{hiroshi0606,ykitamura}@kwansei.ac.jp

Abstract　We have collected a lot of medical parallel texts with translations to support multilingual medical interpretation services for foreign visitors. When we retrieve parallel texts, we face a problem concerning narrowing the search result and ambiguous expression of parallel texts. We propose a parallel text retrieval method based on similarity measure using N-gram, and show its performance. This method provides an advantage that we can utilize the search inputs as cues to collect new parallel texts.

Key words　Similarity Search, N-gram, Information Retrieval, Parallel Text

1. はじめに

国際化の進展に伴い，多言語コミュニケーションの機会がますます多くなっている．日本の病院では，日本語でコミュニケーションのできない外国人が病院を訪れるようになっており，適切な通訳が求められている．一方で，外国語対応が可能な医療機関として公表されているのは，全医療機関のうち，およそ10 %程度であり，またそのほとんどが英語での対応である [1]．そのため，英語以外での対応も求められるが，多言語を習得することは非常に困難であり，機械翻訳の応用が期待される．

近年，翻訳技術は急速に進展しているものの，翻訳精度には限界があり，高精度の翻訳を行うことは非常に困難である [2]．このような正確な通訳が求められる場では，翻訳の品質が保証されている用例の利用に頼らざるをえない．

用例ベースのコミュニケーションに関しては，医療用例対訳を使った対面コミュニケーションを支援する医療受付支援システム M^3 がある [3]．また旅の指差し会話帳 DS(注1)では，画面に言葉を表示することで相手に意思を伝えられる持ち運び可能な携帯端末アプリケーションで，旅先で状況に合った用例対訳を検索することができる．このように用例を利用することで，様々な場面でのコミュニケーションが可能となる．

用例を検索する際，一般的な検索手法は部分一致検索やカテゴリ検索であるが，意図した用例を検索できない場合もある．部分一致検索では，一般に名詞で検索するため，それらが用例文字列に含まれていない場合や，異なる表現で登録されている場合は検索できない．例えば，ユーザは「舌がまわらない」の英訳を取得したくて,「舌」というキーワードを入力しても，登録されている用例が「呂律がまわらない」では，それが検索候補として現れない．一方でカテゴリ検索では，登録されている用例対訳が増加するとともに，それらが属さていれるカテゴリ内での検索も増加するため，用例対訳の増加に比例してユーザ

(注1)：http://www.nintendo.co.jp/ds/aubj/

の検索コストも増加する．このように従来の検索では，検索の絞り込みや用例表現のゆらぎに関して問題がある．

そこで本研究では，N-gram モデルを利用し，検索文に類似する用例を検索する手法を提案する．また本手法を用いた用例検索システムを開発した．本論文では，以下，2 章おいて N-gram を利用した用例間の類似度に基づく検索手法について述べる．3 章おいて本手法を用いて実装した用例検索システムについて述べる．4 章において用例検索システムの評価について述べる．5 章において本手法の医療分野への応用について述べる．6 章において関連研究について述べる．最後に 7 章において本論文の結論について述べる．

2. N-gram を利用した用例間の類似度に基づく検索手法

2.1 用例対訳と検索手法

用例対訳は以下のような単語や文章の原文とその翻訳文の組み合わせである．

- 用例:「健康保険証」
- 対訳:「Medical Insurance Card」

- 用例:「飲み込みにくい」
- 対訳:「difficulty in swallowing」

これらの用例対訳の集合がデータベースに格納され，その検索には部分一致検索やカテゴリ検索が用いられる．しかし，部分一致検索では，一般に名詞や形容詞など短いキーワードを入力することが多く，文章で登録されている用例対訳を検索する場合，ユーザにとって自分が想定している検索要求に対して，適切な検索語を選択することは必ずしも容易ではない．例えば，「吐きたくなったら，知らせて下さい」という用例を部分一致検索する場合,「吐く」や「吐きたい」で検索しても，この用例を検索できない．このように用例に含まれる表現を正確に入力することは困難である．またカテゴリ検索では，用例対訳の増加に比例して用例の選択が面倒になる問題があげられる．そのため，カテゴリ検索は小規模のデータベースに対してしか有効でない．

一方で本研究で提案する検索手法は，検索文字列と類似している用例対訳を検索する，類似検索と呼ぶものである．類似検索は検索文字列に似ている文字列を入力するだけで，検索結果を得られる．またカテゴリ検索と違い，データベースの規模による影響を受け難い．

2.2 N-gram モデルの利点

検索文字列とデータベースに登録されている用例が，類似しているかどうかを判定する必要がある．そこで，検索文字列と用例間の意味合いの近さを表す尺度として，N-gram モデルを用いる．N-gram モデルとは，情報理論の創始者として知られる C.E.Shannon が考え出した，ある文字列の中で N 個の文字列または単語の組み合わせがどの程度出現するかを調査するものである [4]．本手法に N-gram モデルを利用する理由は 3 つある．

- 検索漏れがない

一般的な部分一致検索では，検索文に対して形態素解析が行われている場合が多い．形態素解析は，辞書を使って文字列を，意味を持つ最小の単位 (形態素) に分解する処理である．このため形態素解析で文字を区切ると，意味のある単語に基づいて検索ができる．しかし，用例対訳を検索する場合は，正確な用例表現を入力することが困難であるため，意味のない文字列を拾わないために，検索漏れが生じる可能性がある．そのため本手法では，用例対訳を n 個の文字の並びである N-gram に分割して検索を行うことによって，文章に含まれる単語を無視して文字列単位で認識させ，検索漏れを防ぐ．

- 辞書のメンテナンスが不要

部分一致検索では，辞書を使って文字列を形態素に分解するが，辞書にその形態素が存在しない場合は適切に区切られない．一方で，N-gram では文字の意味は考慮せずに，一定の長さ N で文章を区切って処理を行うため，辞書が必要ない．

- 多言語への展開が容易

用例対訳は多言語にわたって存在する．部分一致検索では，多言語にわたる形態素解析エンジンを必要とする．一方で，N-gram は文法的な解析を行わないため，言語に依存することなく処理できる．

このように用例対訳を検索する場合は，N-gram モデルを利用した検索が適していると考えられるため，本手法ではこれを用いる．

2.3 N-gram に基づく用例間の類似度計算

N-gram では，隣り合った文字列または単語の組み合わせを連続要素と呼ぶ．例として，文字列「健康保険証」の連続要素を以下に示す．

「健康保険証」の連続要素

- 1(uni)-gram「健」「康」「保」「険」「証」
- 2(bi)-gram「健康」「康保」「保険」「険証」
- 3(tri)-gram「健康保」「康保険」「保険証」

本手法では，検索文字列とデータベースに登録されている用例間において，この連続要素がいくつ共起関係にあるかを調べる．日本語の共起関係では 2-gram か 3-gram が適しているため(注2)，本手法には 2-gram を用いる．ここで 2-gram の共起関係を例として示す．2-gram 図 1 のような検索文字列「昨日から頭が痛い」と用例文字列「頭が昨日から痛い」は 5 つの共起 (図中の＊) が存在する．

ここで a を検索文字列の長さ，b を用例文字列の長さとする．a,b の連続要素数は，2-gram のため，それぞれ a-1，b-1 となる．これらの値のうち，大きい方を分母とし，検索文字列との共起関係数を分子とした数値を類似度 N と呼ぶ [5,6].

$$N = \frac{\text{共起数}}{max(a,b) - 1} \tag{1}$$

図 1 の例では，N=5/7 である．この N の値が大きくなるほど，検索文字列と用例文字列が類似しているといえる．

(注2)：http://nlp.nagaokaut.ac.jp/n-gram

で入力すると目的の用例を得られた」との意見が多く出た．類似検索は用例表現のゆらぎを吸収できるため，被験者の多くが 2 回以下の試行回数で目的の用例を取得できた．また「入力が短いほど，目的の用例を取得できなかった」との意見も出た．これらから，類似検索では入力された文字列が文章のように長ければ，目的の用例をより取得しやすいことが分かった．

類似検索において目的の用例を取得し難い場合もあった．それは「診察室はどこですか？」という検索文に対して,「診察室はどちらにありますか？」という用例が登録されているのに，文末表現が同じである「トイレはどこですか？」や「駐車場はどこですか？」という用例が上位の検索結果として表示された場合であった．実験後のアンケートでも,「肯定文と疑問文を分けて検索したい」との意見が出た．疑問文の用例は文末表現が似ているため，類似度は大きいが，意味合いでは近くない用例が上位の検索結果として表示されたと考えられる．

この問題の解決策としては，まず検索文におけるキーワードを見つけて，その連続要素に重みをもたせる必要があると考えられる．まず検索文中からキーワードを見つける手法を考える．転置インデックスにおいて「です」や「か？」といった連続要素は，たくさんの ID を格納している．そのため，これらの連続要素は一般的によく使われる表現であると考えられ，検索文においてキーワードである可能性は低い．一方で「診察室」の連続要素である「診察」と「察室」は,「です」や「か？」に比べ ID の格納数が少ないと考えられ，キーワードである可能性は高い．そのため，これらの連続要素に重みをもたせて類似度計算を行うことで，キーワードを重視した類似度検索が可能になると考えられる．

5. 医療分野への応用

5.1 医療用例対訳収集システム

本研究で提案した検索手法は，様々なドメインの多言語用例対訳検索に利用できると考えられる．とりわけ本研究では，今後の展望として，医療分野への応用を考えている．

国際化の進展に伴い，日本語でコミュニケーションのできない外国人が病院を訪れるようになっているが，急病になった外国人患者を支援している医療通訳ボランティアの人材が少ない．そこで，ボランティアに代わる，翻訳の品質が保証されている用例を使った用例ベースのコミュニケーション支援が有効であると考えられる．しかし，必要な医療用例対訳の数は膨大であり，既存の医療用例対訳のみでは不十分であるため，その蓄積には用例を投入する医療従事者や翻訳をする通訳者など，多くの人々が関与する必要がある．そこで図 6 に示す，多くの人々が Web 上で協調して，翻訳の品質が保証されている医療用例対訳を収集する医療用例対訳収集システム TackPad が開発されている [7-9].

5.2 本手法の応用例

医療用例対訳収集システムでは，ユーザは多様な用例対訳を十分な数だけ思いつくことが簡単ではないため，登録するには手間がかかる．そこで，用例対訳収集システムに本検索手法を取り入れる．

図 6 医療用例対訳収集システム Tackpad
Fig. 6 Medical parallel texts collecting system Tackpad

その手法としては，データベースに登録されていない検索文を新たな用例として自動取得することが考えられる．まずユーザは，本研究で用いた検索手法である，類似検索を使った用例対訳検索システムを使って，用例対訳を検索する．もしユーザが目的の用例対訳を得られなかった場合，類似検索を用いているため，検索文をデータベースに登録されていない新たな用例として獲得できる．図 7 では，検索文「歯茎から出欠する」で類似検索を行ったが，類似用例は存在しなかった．しかし，検索文は新規用例として獲得できる．

図 7 新規用例の獲得
Fig. 7 Collecting new parallel texts

このように，本研究で用いた類似検索は，ユーザが検索に失敗した際に入力されていた検索文を新たな用例として自動収集する手法として応用できると考えられる．前章の評価実験において，類似検索を行うシステムのログを解析した結果，ユーザはいくつかの文章を入力することが分かった．検索文の中には「息子を診察してください」,「調子はどうですか？」,「子供がすごい熱です」といった，データベースに登録されていない文章を得ることができた．これらの文章を新規用例として利用することにより，医療用例対訳収集システムに用例対訳を登録する手間が省けると考えられる．

5.3 医療用例対訳の収集と利用

本システム，医療用例対訳収集システム Tackpad，病院において患者の受付を支援する医療受付支援システム M^3 における医療用例対訳の収集と利用の全体像を図 8 に示す．医療従事者や通訳者は，医療用例対訳収集システム TackPad に医療用例対訳を登録する．しかし，十分な数の医療用例対訳を登録するには手間がかかる．そこで，本手法を用いて用例対訳を登録す

る負担を軽減させる．まず医療従事者や通訳者は本システムを使って，知りたい用例対訳を検索したり，これから登録したい用例対訳が，類似した用例も含めて既に TackPad に存在しないかどうかをチェックしたりする．これらの検索の結果，類似度が小さい検索文の中から，新たな用例として採用できるものを抽出する．抽出した検索文を TackPad に登録することにより，用例対訳が手間なく蓄積できる．蓄積された医療用例対訳は，医療受付支援システム M^3 によって医療従事者と外国人患者との対面コミュニケーションに使われる．

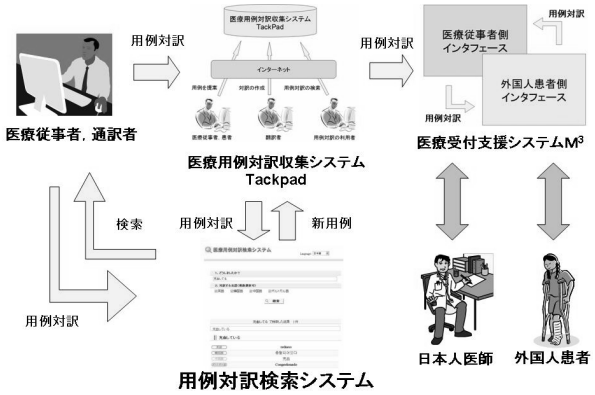

図 8　医療用例対訳の収集と利用
Fig. 8　Collecting and using medical parallel texts

6. 関 連 研 究

入力された検索文に類似した文章を検索する研究を紹介する．

著作権侵害文章の抽出を目的として，既存の商用検索エンジンを用いた類似文章検索システムに EPCI がある [10]．EPCI では，検索エンジンを用いて候補ページ集合を集める，候補ページと入力文章との類似度に基づき候補ページ集合をランキングする．候補ページの収集ステップでは，入力文章から文節を単位とした N-gram をクエリ集合として生成し，各クエリについて候補ページ集合の平均類似度が閾値以下になるランキングまで上位から取得を行うことで，網羅性の向上を実現する．

また，話し言葉である発話を対象として機械翻訳を行う研究がある [11]．話し言葉はその特有の性質が一因となって，機械翻訳を行った際，適切な翻訳文が得られない場合がある．そこで適切な翻訳文が得られなかった場合に，類似文検索技術を用いることで適切な訳文を得る手法がある．機械翻訳に与えた入力文が翻訳不能文と判明した場合, 翻訳可能文コーパスからその入力文に対する類似文を検索する．類似文が検索できた場合は，入力文を類似文と置き換え翻訳を行う．検索対象となる候補文と入力文の間の類似度は, 候補文と入力文の間で共通する N-gram の比率に基づいて算出する．これにより，翻訳不能である入力文に対しても，翻訳可能な類似文を検索することで, 適切な訳文を得ることができる．

また，入力文と正解文の類似度を測る指標として BLEU がよく知られている [6]．入力文と正解文の N-gram の重なりによって類似度を判定している．本手法で述べた類似検索は，BLEU で用いられている N-gram の定義式を参考にしている．入力文と正解文の N-gram の共起数を類似度とした場合，正解文よりも入力文が長い場合は，正解文にある N-gram 要素数を越えないので，類似度は小さくなる．しかし，短い場合は適合率が高くなりやすいため，類似度は大きくなる．そこで BLUE では，正解文より短い入力文に対してのみ，文長に応じたペナルティを課して補正を行う．これに対して本手法は，入力文と用例文において，文長が大きい方を基として小さい方との適合率を求めている点で異なる．

7. まとめと今後の課題

本研究では，ユーザが入力した検索文に類似した用例を検索できる用例対訳検索システムの開発を目指した．またその検索手法としては，N-gram モデルを使った類似度計算を用いた．これにより，ユーザは入力文に近い意味合いの用例を検索することが可能となった．

今後の課題は 2 つある．1 つ目は，検索文と文末表現が同じ用例が検索結果の上位にくる問題がある．これに対しては，検索文字列におけるキーワードを抽出し，その連続要素に重みをもたせて類似度計算を行うことで，意味合いの近い用例との類似度を上げることができると考えられる．2 つ目は，本検索手法は特定分野に限らず利用できるため，実際に色々な分野で本手法を使ってもらい，その利用法を調査することである．

謝辞　本研究の成果は戦略的情報通信研究開発推進制度 (SCOPE) に基づくものである．

文　　献

[1] 財団法人国際コミュニケーション基金: 在日外国人医療におけるコミュニケーションギャップの現状調査と改善策の研究, http://www.icf.or.jp/icf/out/download/Foreigner_medical_treatment.pdf.
[2] Toru Ishida: Language Grid:An Infrastructure for InterculturalCollaboration, SAINT-06, pp.96-100, 2006.
[3] 宮部，吉野孝他:異文化間コミュニケーションのための用例を用いた医療受付対話支援システムの開発.DICOMO2006.
[4] Claude Shannon: A Mathematical Theory of Communication, Bell System Technical Journal, vol.27, pp.379-423 and 623-656, 1948.
[5] 内山将夫，コーパスベースの機械翻訳, http://www2.nict.go.jp/x/x161/members/mutiyama/saitama-u.html
[6] Kishore Papineni, Salim Roukos, Todd Ward and Wei-Jing Zhu:BLEU: a Method for Automatic Evaluation of Machine Translation. ACL pp.311-318，2002.
[7] 岩部正明，村上陽平，重野亜久里，石田亨:Web サービス連携を用いた医療用例対訳の収集と利用，電子情報通信学会技術研究報告，AI2006-28，pp.17-22，2006.
[8] 岩部正明，田淵裕章，村上陽平，重野亜久里，石田　亨，北村泰彦，河原達也，吉野　孝，津村　宏:言語グリッドを用いた医療用例対訳の収集，第 69 回情報処理学会全国大会，デ-11, 2007.
[9] 福島拓，吉野孝，医療分野を対象とした多言語用例収集 Web システム TackPad の開発，http://www.wakayama-u.ac.jp/ yoshino/lab/research/fukushima/
[10] 下畑光夫，隅田英一郎，松本裕治:発話を対象とした類似文検索と機械翻訳への適用. 自然言語処理，Vol.11，No. 4，2004.
[11] 田代 崇，上田 高徳，平手 勇宇，山名 早人:検索エンジンを用いた類似文章検索システム EPCI の評価.DEWS2008,2008.