機械学習を用いた語源的英単語分割手法の開発
― 同語源単語を比較する学習アプリケーション ―

1　背景

外国語単語学習は，しばしば任意の文字列の羅列に対して無関係な意味を対応づけるという苦痛な作業になりがちである．このような場合に語源情報を生かして学習すると良い，ということはしばしば言及される．外国語単語学習のために語源情報を整理した書籍や Web サイトは既に複数存在する．しかしながら，このような情報資源は手作業でまとめられているため，中級以上の単語や専門用語に関してはカバーできていない．一方で語源情報を機械的に解析してデータベース化した取り組みはいくつかあるが，必ずしも学習者が使いやすいようになっていない．

2　目的

本プロジェクトの目的は，語源情報を機械的に整理し，それを外国語単語学習に向けて使い勝手の良い学習アプリケーションとして提供することである．

英単語を語源とともに暗記する利点は，共通の祖先を持つ単語類（同語源単語）を一括りに暗記できることである．同語源単語は意味とスペリングの両面で先祖となる単語の特徴を引き継いでいるため，同語源単語内ではそれらに共通性があることが多い．したがって，英単語を確認する都度そのグループの意味とスペリングの特徴を把握しておけば，知らないスペリングを見た時に意味を推測できるようになる感覚が身に付く．このような学習を手軽に行えるようにするため，本プロジェクトで作成する学習アプリケーションで提供する機能は，検索単語のスペリングに対して語源を紐づけること，及び，紐付けた語源を共通にもつ他の単語を並べて表示し，比較できるようにすることである．

また，作成する学習アプリケーションの主要なターゲットは中学や高校の基礎レベルといった，すでに手作業ベースの処理で対応できている層ではなく，大学レベルから専門用語といった中級以上に分類されるような英単語学習である．また，英語以外の外国語にも利用できることも重要である．従って本プロジェクトの解析手法は機械的であって，情報源の拡張に対応できる方法である．

3　開発の内容

本プロジェクトでは，まず，語源情報を機械的に扱いやすい，グラフ形式にまとめる（以下，このようにまとめたものを「語源ネットワーク」と呼ぶ）．次に語源とスペリングを紐づけるために，このデータを用いて，任意の単語に対してそのスペリングを語源的意味のある最小単位と対応する語源語のリストに変換する仕組みを開発する．また，複数の同語源単語の語源を比較できるようにするため，複数単語の語源ネットワークを見通しよく描画する手法を開発し，Web アプリケーションを作成する．

3.1　高品質な語源ネットワークの構築

Web 上で利用可能な散文形式の語源の記述を取得して，グラフ形式にまとめた語源ネットワークを作成した．これは，スペリングと，所属する言語，id の情報を持つノードに対し，継承関係のある場合に先祖から子孫の方向に有向エッジを持たせたものである．高品質な語源ネットワー

クを構築するため，冗長性という観点で取り組んだ．複数のデータソースを解析し，解析結果を統合することで，1 つのエッジに対してその根拠が複数のデータソースにあるようにすれば，誤解析の影響を抑えられるからである．

データソースには，複数言語版の Wiktionary を Wiktextract で事前変換したものや，スクレイピングアルゴリズムを記述して収集した Web データを用いた．収集した語源の記述から，Transformer モデルを用いて継承関係を抽出した．英語版 Wiktionary を独自にルールベースで解析・スクリーニングしたデータセットの一部を用いてモデルをファインチューニングし，その他のデータの解析を行う方針を採った．プロンプトは，散文形式の語源の記述から，その中に含まれる単語に対しその先祖は何かを問う形とした．また，回答の正確性を上げるため，Few-shot 形式の問い方にした．先祖がない場合は，空白を返すようにした．

Transformer モデルには学習済みの LLAMA3 あるいは T5-base を試してみた．LLAMA に関しては推論が遅いことや，入力した語源の文章にない単語を出力してしまう挙動が多かった．また，今回のような単語抽出のタスクには，Decoder-only モデルである LLAMA よりも Encoder-Decoder モデルである T5 の方が適していると判断した．T5 モデルの推論に関しては，FasterTransformer を用いて約 4 倍に高速化した．また，地の文が英語以外の場合は MarianMT モデルで翻訳してから解析した．この翻訳処理は ONNX 形式にエクスポートしたり T5 の推論処理と非同期に行うようにして高速化した．

このようにして取得した継承関係のデータを統合して一つのネットワークにまとめる．この際様々な種類の表記揺れが問題となる．活用形に関しては Wiktionary の活用表から活用関係のグラフデータを作成し，その結合成分をとったりクラスタリングをしたりして，活用のグループを作り表記揺れを統一した．同形異義語に関しては，SentenceTransformer で意味表現の埋め込みベクトルを作成し，その内積を比較することで分類した (Figure 1)．また言語名の表記揺れを手作業で直したり，継承関係の粒度を語源ネットワークのショートカットを修正することで統一したりした．

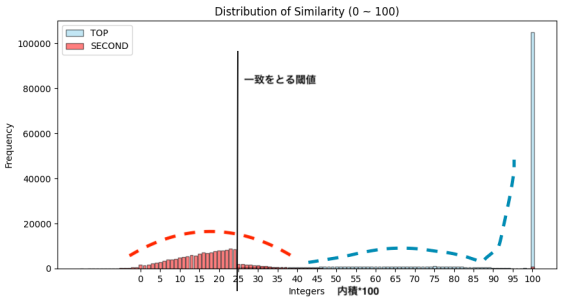


Figure 1: 意味表現のベクトルの類似度の分布．同じスペリングの単語の異なる意味に対して内積を取った時，その値の分布を表したものである．横軸は $0 \sim 1$ の内積の値を 100 倍にしており，bin 幅 1 のヒストグラムにしている．青色が最も一致したもの，赤色が 2 番目に一致したものの分布であり，この 2 つの分布の山が区別できているので同形異議語を区別できている状況証拠となる．

最終的には語源ネットワークに含まれるエッジのうち 32.8% に関して，重複性を確保できた状

態となった．

3.2　語源的英単語分割の開発

スペリングに語源との相関付けを行う技術を開発する．主に Cross-Attention レイヤを直接見ることで，アラインメント情報を取り出す手段を取ることにした．さまざまなモデル構造を試した結果，Encoder および Decoder の Embedding 層のすぐ後に，LSTM を導入した構造が扱いやすかった（Figure 2）．Decoder には子孫の一単語，Encoder にはその先祖（1 つあるいは複数）を入力し，Decoder の出力が Decoder の入力の次文予測となるように学習させた．データセットは子孫と，語源ネットワークを参照して取得した先祖のペアのリストとした．トークナイザは文字ごとに数値を割り当てるものとして行った．この結果，Cross-Attention の最終層を取り出すと，人間の直感とも整合するヒートマップが得られた (Figure 4)．

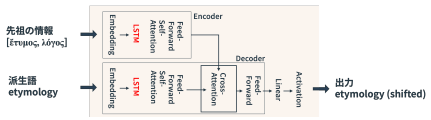
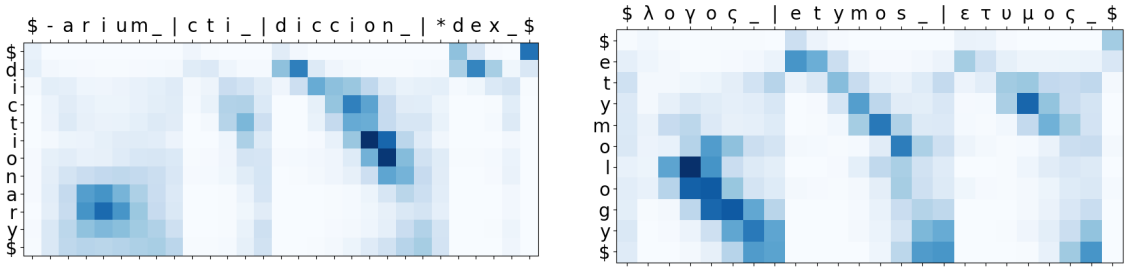


Figure 2: モデルの構造．赤字の LSTM が新たに入れた層である．

<st>egg-shaped<en>	<st>egg-shaped[en]<en>
<st>intrasegmentally<en>	<st>-ly[en]|intra[la]|segmental[en]<en>
<st>wikipedically<en>	<st>encyclopedia[en]|algebraic[en]|-ik[enm]|wiki[en]|-al[en]<en>
<st>pharmacodependence<en>	<st>-ence[en]|dependance[fro]|φαρμακον[grc]|depend[en]<en>
<st>noble<en>	<st>-ibilis[la]|*gneh$_3$-[ine-pro]<en>
<st>earthshine<en>	<st>*erþu[gmw-pro]|*skinan[gmw-pro]<en>
<st>guardhouse<en>	<st>guarde[fro]|hous[enm]<en>
<st>departmentwide<en>	<st>-wide[en]|departement[fro]|-ment[fro]|de-[fro]|departire[LL.]|partir[fro]<en>
<st>unpollutable<en>	<st>por-[la]|-luo[la]|por-[la]|*lew-[ine-pro]<en>
<st>troll<en>	<st>trollen[enm]<en>
<st>bookdealer<en>	<st>*boko[gem-pro]|dælere[ang]<en>

Figure 3: 学習に用いたデータセットの一部



“dictionary”　　“etymology”

Figure 4: 語源とのアラインメントの例．なお，アンダーバーのところには先祖の言語コードに対応するトークンが埋め込まれている．また，ドルマークは文字列の始まりや終わりを表すスペシャルトークンである．

最後に，語源的分割を作成する．語源ネットワークの派生語に対してその先祖の部分を展開し，いくつか代表的な先祖を選択する．選んだ先祖に対して学習済みのモデルで推論を行い，Cross-Attention の最終層を抽出し 1 次元に平坦化する．こうして得られたスペクトルに対し，語源グラフを用いたクラスタリングと，スペクトルを用いたクラスタリングを組み合わせて調整し，主要な先祖を選択することで一意な語源的分割を定めた．

3.3　逐次更新可能な階層的グラフ描画

同語源単語を並べて表示するために，ノードやエッジを後から追加できる，すなわち逐次更新可能なグラフ描画が必要である．一方で語源ネットワークは時系列のあるデータなので階層的な

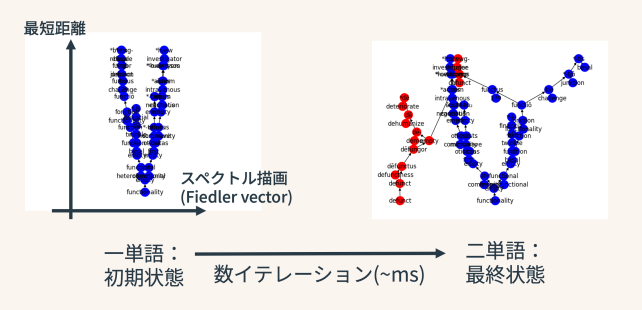


Figure 5: グラフ描画のアップデートの様子．青が初期ノード，赤が新たに追加したノードである．初期ノードに対して階層的スペクトル描画をしたものが左の図であり，そこからノードを追加し，数回のイテレーションを回した際に得られた描画が右の図である．

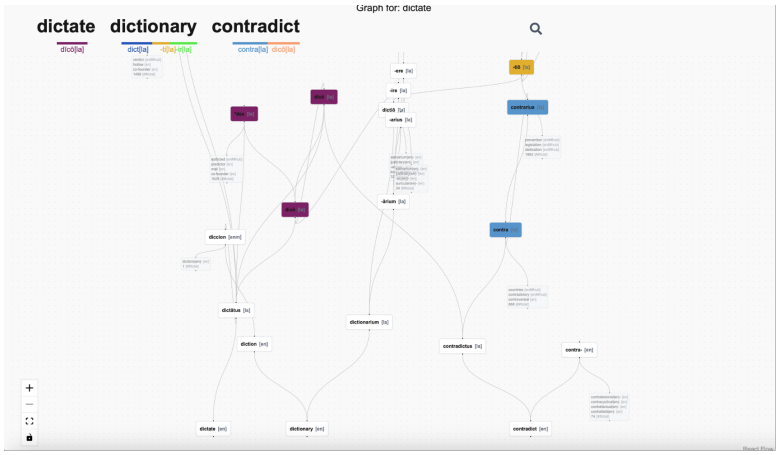


Figure 6: Web アプリケーションで語源情報を検索した画面の例

描画でないと見通しの悪い表現になってしまう．

そこで，縦軸方向は階層番号で固定しつつも，横軸方向に 1 次元のスペクトル描画を用いる描画法を採用した．これによって，逐次更新が可能であるほか，クラスタを強調できるので見通しの良い描画方法となる．実際 Figure 5 のように，数回のイテレーションを回すだけで，階層性やクラスタ性を保ちながら 2 つのグラフを融合することができている．

3.4　学習アプリケーションの作成

以上の語源ネットワーク，語源的単語分割，グラフ描画の技術を集合して，英単語学習のための Web アプリケーションを作成した．Figure 6 はこの学習アプリケーションで “dictate”, “dictionary”, “contradict” という語源情報を並べて表示した際の画面である．左上に表示されているのが語源的分割で，そのアンダーラインの色と対応するグラフ上のノードの色が同じになるようにしている．また，共通部分が取れた先祖もまた同じ色で着色するよう工夫している．複数単語の共通祖先のノードは赤く縁取って強調するようにしている．このような機能によって直感的に語源やスペリングの共通性を理解できるようになっている．

3.5　語源的新語生成

語源的英単語分割で作成した語源分割モデルの Encoder には実際に関わり合った先祖を入力して子孫を推定していたが，歴史上は交わり合わなかった 2 つの先祖を入力することも可能である．この際に Decoder の出力を見れば語源的に自然な新語が生成されていると考えられる．

そこで，サービス名等の新語生成時に複数の単語を合体させて一つの単語を作ることに資することを目的として，入力した 2, 3 の英単語に対して，語源ネットワークでその先祖をたどり，いくつかサンプリングして語源分割モデルに入力し，新語を生成するサービスを作成した．

4　従来の技術（または機能）との相違

単語学習のサービスとしては，語源情報を用いた単語学習を中級以上の英単語や専門単語，また英語以外の外国語でも行えるようになったという点が新しい．また，同語源単語を比較するという構想に基づいたかつてない単語学習のサービスである．語源的分割を表示することで同語源単語のスペリングの共通点を把握できる機能も今までにないものである．

言語処理の研究分野では語源ネットワークを構築する研究がいくつかなされてきたが複数データセットを大規模に統一することによって品質を向上する取り組みは新しいものである．また，英語版以外のデータセットも用いることで英語話者がアクセスしにくい言語の情報も取り込めていると考えられ，文化的・産業的にも価値がある．

語源的分割手法としては，機械学習モデルの内部レイヤを，機械的に処理して利用するという新しい発想に基づいている．また，グラフ描画手法としては，一次元でスペクトル描画を用いた例はみられず，本手法は優れた階層性・クラスタ性と逐次更新可能性を両立した応用性の広い描画手法である．

5　普及（または活用）の見通し

本プロジェクトで開発した Web アプリケーションは，今までに語源情報が利用しにくかった水準の単語学習について手軽に利用できる構成になっている．本アプリケーションはそうした学習者への普及や，語源情報を日常的に利活用する分野自体の価値を向上することが期待される．

また，言語学の研究面での活用も考えられる．開発した語源ネットワークはかつてない多種のデータセットに基づくものであるため，品質や多様性の面で優れていると考えれる．今後整理・評価をした上で，同分野の研究に利用可能な資源として使いやすいように整える．また，語源的分割を作成するために異種言語の単語間でのアラインメントを取ることは，それら言語の関わりを分析する手法としての広がりが考えられる．このような研究は，多様な言語の文化的理解と共に，言語処理の産業を多言語に拡大する上でも重要な技術である．

6　クリエータ名（所属）

中澤 正樹（東京大学 大学院理学系研究科 物理学専攻）

（参考）関連 URL

`https://etymore.com`