IPA

2024 年度 未踏 IT 人材発掘・育成事業 採択案件評価書

1. 担当 PM

五十嵐 悠紀（お茶の水女子大学 理学部 情報科学科 准教授）

2. クリエータ氏名

中澤 正樹（東京大学 大学院理学系研究科 物理学専攻）

3. 委託金支払額

2,880,000 円

4. テーマ名

機械学習を用いた語源的英単語分割手法の開発

5. 関連 Web サイト

https://etymore.com

6. テーマ概要

本プロジェクトでは英単語の語源情報を機械的に整理し、外国語単語学習を行いやすいアプリケーションを開発した。語源情報を機械的に扱いやすいグラフ形式「語源ネットワーク」にまとめ、語源とスペリングとの紐づけを行った。それを複数の同語源単語の語源を比較できるように、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 アプリケーションとして一般の人が使える形にまとめた。

7. 採択理由

本プロジェクトは、英単語を対象として語源的に意味のある最小単位に分割する機械学習モデルを開発する提案である。英単語を語源で着目しなおし、一般ユーザにわかりやすくそれを提示することで、単語を知らなくてもその意味を推測できるセンスが積めるようになる支援にもつながるアプリケーションも開発する。すでに語源データの解析部分を開発し、ファインチューニングのプロトタイプもできていたことで、プロジェクト期間中に開発するシステムの技術的側面や乗り越えるべき課題も明確であったことと、本手法によりユーザの英語学習の幅が広がりそうなことに期待して採択とした。英語だけでなく、語

学全般に使える技術に展開できる可能性にも期待する。

8. 開発目標

本プロジェクトでは、

(1) 機械学習を用いて任意の英単語を語源的に意味のある最小単位に分割する方法を開発すること
(2) 分割した単語を学習者が見やすい形に描画すること
(3) それらを単一のアプリケーションとして使えるようにすること

を目標とした。

9. 進捗概要

本プロジェクトでは、語源情報を機械的に扱いやすい、グラフ形式にまとめた「語源ネットワーク」を構築した。これは、スペリング、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。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などして語源ネットワークを構築し、高速化も行った。

次に語源とスペリングを紐づけるために、語源ネットワークのデータを用いて、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した。スペリングに語源との相関付けを行う技術として、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した。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（図 1）。

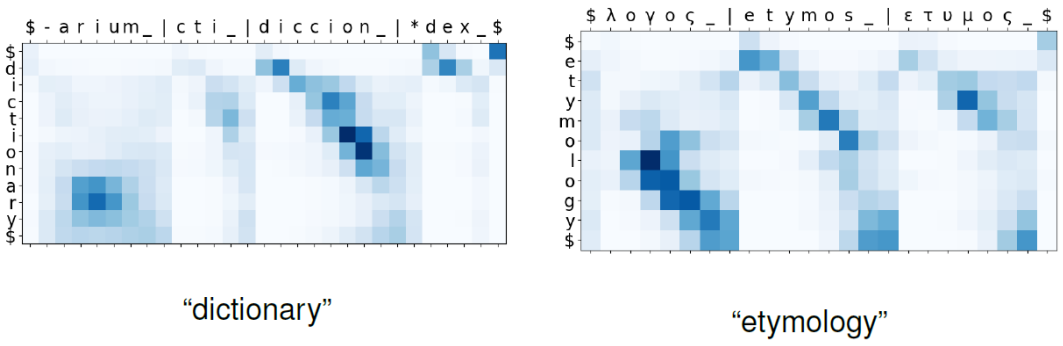

図 1：語源とのアラインメントの例
アンダーバーのところには先祖の言語コードに対応するトークンが埋め込まれている。
また、ドルマークは文字列の始まりや終わりを表すスペシャルトークンである。

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グラフ描画する手法を開発した。そして、語源ネットワーク、語源的単語分割、グラフ描画の技術を利用して、英単語学習のための Web アプリケーションを作成した（図 2）。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。左上に語源的分割が表示されており、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っている。また、共通部分が取れた先祖も同色で表示されている。複数単語の共通祖先のノードは赤く縁どって強調するようにしている。このような機能によって学習者が直感的に語源やスペリングの共通性を理解することができるように設計した。

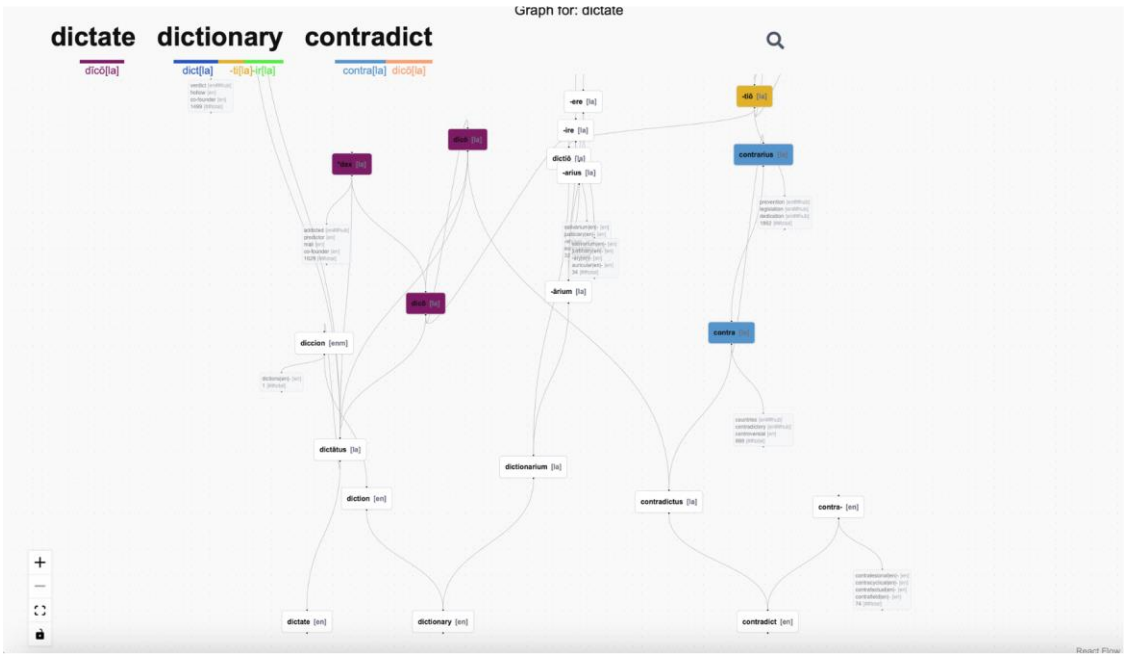

図 2：Web アプリケーションで語源情報を検索した画面の例

また、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とを利用し、複数の単語を合体させて一つの単語を作ることができる「語源的新語生成」機能も開発した。サービス名等の新語生成時に便利である。

プロジェクトを進めるにあたっては語源的分割手法とグラフ描画手法の開発が大きなカギであり、それぞれの手法考案にかけるバランスなどに苦労しながら進めてきた。結果として、バランスの良い、Web アプリケーションとしてまとめて一般に公開することができた。

10. プロジェクト評価

当初目標としていた 3 点、

(1) 機械学習を用いて任意の英単語を語源的に意味のある最小単位に分割する方法の開発
(2) 分割した単語を学習者が見やすい形に描画すること
(3) それらを単一のアプリケーションとして使えるようにすること

について丁寧に実装を進めた結果、すべて達成し、Web アプリケーションとして一般に公開することができたことは評価に値する。特に語源分割手法とグラフ描画については、技術的には非常に面白い成果と評価する。

一方で、語源分割に関して言語学的な正しさの検証やユーザインタビューまではたどり着けず、有用性を示す部分で PM の期待には届かなかった。グラフ描画についても、対象ユーザがどのようにこの機能を使って探索をするのか、この機能によって実際に英語学習を支援することができるのか、といったユーザビリティの部分を、ユーザインタビューを重ねてもう少し検証して欲しかった。

11. 今後の課題

現状は中澤氏自身が利用して使ってみた状態に留まっているため、一般公開したこともきっかけとして、多くのユーザに使ってもらい、フィードバックを得た上でどのように使われているのか検証してみて欲しい。加えて、本プロジェクトの提案当初に予測していた通り、ユーザがこの Web アプリケーションを利用し語源情報を生かして英単語の学習をすることで、知らないスペリングを見た時に意味を推測できる感覚が身に付くのか、ユーザインタビューなどを進めて検証していって欲しい。