EtymoLink: A Structured English Etymology Dataset

Yuan Gao, Weiwei Sun

Department of Computer Science and Technology, University of Cambridge

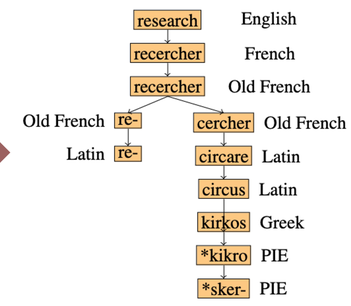
UNIVERSITY OF CAMBRIDGE

Motivation

Most etymological data lives in scholarly articles, etymological dictionaries, or web resources.

- Such formats are inherently unstructured and not suited for computational approaches.
- Transferable model to structure other lexicographical resources.

research, (*v.*)
1590s, "investigate or study (a matter) closely, search or examine with continued care," from French recercher, from Old French recercher "seek out, search closely," from re-, here perhaps an intensive prefix (see re-), + cercher "to seek for," from Latin circare "go about, wander, traverse," in Late Latin "to wander hither and thither," from circus "circle" (see circus).
The intransitive meaning "make researches" is by 1781. Sometimes 17c. also "to seek (a woman) in love or marriage." Related: Researched; researching.



Abstract

This paper presents a methodology and empirical study for creating a **structured etymological dataset** suitable for computational analysis.

- **Data Source:** Etymonline.
- **Method:** Manual annotation and fine-tuning the FLAN-T5-base model.
  - **0.902** BLEU Score
- **Dataset:** Over 103,000 relationships covering 63,603 English lexical terms.
  - **Ground Truth:** 5,361 entries manually annotated.
  - 58,242 automatically annotated.
- **Result:** High accuracy in identifying lexical terms; rooms for improvement in identifying the source language.
  - A plateau effect illustrating the model can effectively structure data with limited annotations.

Methodology

Data Collection

- 63,603 entries were extracted from Etymonline.
- Differentiated homographs, preserved hyperlinks

research *(v.)*
1590s, "investigate or study (a matter) closely, search or examine with continued care," from French recercher, from Old French recercher "seek out, search closely," from re-, here perhaps an intensive prefix (see re-), + cercher "to seek for," from Latin circare "go about, wander, traverse," in Late Latin "to wander hither and thither," from circus "circle" (see circus).
...

Manual Annotation

- 5,361 entries manually annotated.
- Ensured diversity in word initials.

research *(v.)*
research_E, recercher_F
recercher_F, recercher_OF
recercher_OF, re-_OF, cercher_OF
cercher_OF, circare_L
circare_L, circus_L

- Each line represents an descendency relationship, i.e. edge(s) in the graph.
- Each term makes up of the word root and the language, separated by an underscore.

Training

- Candidate words are extracted using Regex.
- FLAN-T5-base, an encoder-decoder language model, was fine-tuned.

###INSTRUCTION:extracting etymological relations from text and structuring this information into an edge adjacency list.
###WORD: research (v.)
###TEXT: 1590s, from Middle French recercher, from Old French recercher ""seek out, search closely,"" from re-, intensive prefix (see re-), + cercher ""to seek for,"" from Latin circare ""go about, wander, traverse,"" in Late Latin ""to wander hither and thither,"" from circus ""circle"" (see circus). Related: Researched; researching.
###CAND: recercher, recercher, re-, re-, cercher, circare, circus, circus, Researched, researching"

Evaluation

- The string-based evaluation focuses on measuring the textual similarity between the model-generated output and the target (manually annotated) output.

String-based Evaluation	
BLEU	0.902
Rouge	0.920
ChrF	0.929

Table 1. String-based evaluation results on a held-out dataset of 805 terms

- The edge-based evaluation assesses the structural and relational accuracy of the outputs.

Edge-based Evaluation	
Edge Recall	0.905
Language Label Detection	0.990
Language Label Accuracy	0.909
Word Root Accuracy	0.905
Word Root Levenshtein Distance	0.321

Table 2. Edge-based evaluation metrics. Edge recall is the proportion of the etymological relationships (edges) in the data that the model identified, accurately or not. Language label detection reports the proportion of word roots that received a language label, accurately or not, while language label accuracy reports the proportion of word roots with the correct language label. Word root accuracy reports the proportion of the word roots correctly extracted and word root Levenshtein distance reports the average edit distance of predicted word roots from the actual roots.

Effects of Training Data Size

- One of the motivation of this project is to investigate the feasibility of leveraging LLMs to extract structured data from dictionaries.
- The FLAN-T5-base model was trained with different subsets of the training corpus, including sizes of 1000, 2000, 3000, 4000, and the entire corpus of 4556.

Figure 1. Performance on BLEU, ROUGE, ChrF, Root Accuracy, and Language Label Accuracy over different training data size.

Acknowledgement

We want to thank Li Liang for the data collection and annotation, and Junjie Cao for the initial implementation of a ranking system. We would like to also thank the anonymous reviewers for the insightful and valuable suggestions. This work is partly supported by Cambridge University Press & Assessment.